3D-TV USING DEPTH-IMAGE-BASED RENDERING (DIBR)

CHRISTOPH FEHN
Image Processing Department
Fraunhofer Institute for Telecommunications, Heinrich-Hertz Institut
Einsteinufer 37, 10587 Berlin, Germany
tel.: +49 – (0)30 31002611, fax: +49 – (0)30 3927200
email: christoph.fehn@hhi.fhg.de

ABSTRACT

This paper will present details of a system that allows for an evolutionary introduction of depth perception into the existing 2D digital TV framework [16]. In contrast to former proposals, which often relied on the basic concept of an end-to-end stereoscopic video chain, this new idea is based on a more flexible joint transmission of monoscopic color video and associated per-pixel depth information. From this 3D data representation format, one or more “virtual” views of a 3D scene can then be synthesized in real-time at the receiver side by means of so-called depth-image-based rendering (DIBR) techniques [11, 12]. This paper (a) highlights the advantages of this new approach on 3D-TV; (b) describes a method for the synthesis of stereoscopic images using the concept of a “virtual” stereo camera; (c) investigates the efficient compression of 3D imagery by means of state-of-the-art MPEG coding standards and (d) derives an optimal algorithm for the quantization of 3D scene depth.

1 Introduction

In recent years, joint work in the historically separated research areas computer vision (CV) and computer graphics (CG) has led to the development of novel image-based modeling (IBM) and -rendering (IBR) techniques, which allow to create high-quality “virtual” views of real-world scenes from one or more existing color images and, depending on the particular technology, more or less geometrical information [18]. Such methods are well suited for the implementation of a new, backwards-compatible approach to broadcast 3D-TV. Within the European IST research project ATTEST [16], a system has been developed, in which stereoscopic images are synthesized in real-time from a 3D data representation format consisting of monoscopic color video and associated 8-bit per-pixel depth information (see Fig. 1). This data can be transmitted in a backwards-compatible way over the existing digital broadcast TV infrastructure (DVB-C/S/T) by using MPEG-2 for the compression of the conventional 2D video and any of the more efficient additions to the MPEG family of standards (MPEG-4 Visual or H.264) for the encoding of the additional depth-images.

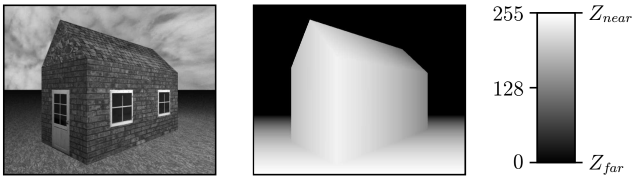


Figure 1: 3D data representation format consisting of regular 2D color video in European digital TV format and accompanying 8-bit depth-images with the same spatio-temporal resolution.

Compared to an end-to-end stereoscopic video chain, the proposed 3D-TV concept has a number of advantages with the most important virtues being the following [3]:

- Backwards-compatibility to today’s 2D digital TV.
- 3D effect can be customized to suit different displays.
- Very low storage and transmission overhead ($< 20\%$).
- Data representation supports multiview 3D displays.
- Allows viewer control over depth reproduction [14].
- Content can be created from existing monoscopic color video using “structure-from-motion” techniques [2, 6].

2 “Virtual” Stereo Camera

The synthesis of stereoscopic images requires the definition of two “virtual” cameras – one for the left-eye and one for the right-eye. With respect to original (reference) view, these cameras are symmetrically displaced by half the *inter-axial distance* t_c (see Fig. 2). To establish the zero parallax setting (ZPS), i.e. to choose the *convergence distance* Z_c in the 3D scene, the CCD sensors of the parallel positioned cameras are translated by a small shift h relative to the position of the lenses. In real stereo cameras, this *shift-sensor* concept is usually prefered over the ‘toed-in’ approach because it does not introduce keystone distortions and depth-plane curvature in the stereoscopic imagery [10, 19]. When implemented with depth-image-based rendering (DIBR), it has the additional advantage that all the required signal processing (e.g. the antialiased resampling of the "virtual" left- and right-eye views) is purely one-dimensional.

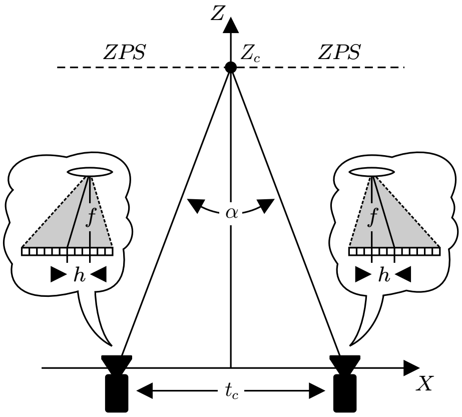


Figure 2: "Virtual" stereo camera using the shift-sensor concept to establish the zero-parallax setting (ZPS).

The transformations that define the new locations (u', v), resp. (u'', v) of the original image points (u, v) in the "virtual" left- and right-eye views can be derived from the theory of *3D image warping* [11, 12]. They result to:

$$u' = \frac{\alpha_u t_c}{2}\left(\frac{1}{Z} - \frac{1}{Z_c}\right) \text{ , resp. } u'' = \frac{\alpha_u t_c}{2}\left(\frac{1}{Z_c} - \frac{1}{Z}\right) \ , \tag{1}$$

where α_u is the focal length of the reference camera expressed in multiples of the pixel width w [20]. (A detailed derivation of these equations as well as a comparison of different approaches for dealing with the inherent problem of disocclusion artifacts can be found in [3, 4].)

Table 1 shows, how the 3D reproduction that results from the application of Eq. (1) 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 original camera and the convergence distance Z_c. The respective changes in parallax, perceived depth and object size are qualitatively equal to what happens in a real stereo camera when these system parameters are manually adjusted. For example, an amplification of the interaxial distance t_c leads to an increase in parallax. This, in turn, leads to an increase in perceived depth without, however, affecting the perceived object size. A magnification of the camera lenses' focal length f, on the other hand side, not only leads to an increase in parallax and perceived depth, it also increases the perceived object size. Finally, changing the convergence to a larger distance Z_c decreases the parallax without affecting the overall perceived depth and object size. However, the 3D scene will appear to be shifted forward on any stereoscopic- or autostereoscopic 3D display.

Parameter	+/–	Parallax	Perc. depth	Obj. size
t_c	+	Increase	Increase	Constant
	–	Decrease	Decrease	Constant
f	+	Increase	Increase	Increase
	–	Decrease	Decrease	Decrease
Z_c	+	Decrease	Shift (fore)	Constant
	–	Increase	Shift (aft)	Constant

Table 1: Qualitative changes in parallax, perceived depth and object size when varying the interaxial distance t_c, the focal length f or the convergence distance Z_c of a "virtual" stereo camera (after [13]).

The main feature that distinguishes a "virtual" stereo camera from a real one is the fact that the stereoscopic system parameters do not longer have to be defined at *capture time*. Rather they can be optimized at *display time* to customize the resulting 3D reproduction for a specific viewing condition. This also allows the user him/herself to adjust the depth percept to meet his/her personal preferences [3].

The effect of interactively changing the parallax:

$$P(Z) = \alpha_u t_c \left(\frac{1}{Z_c} - \frac{1}{Z}\right) \ , \tag{2}$$

by means of a variation of the interaxial distance t_c is also visualized in Fig. 3. The magnified clippings from two "virtual" stereoscopic images show that for $t_c = 24$ mm only a rather low amount of parallax – and thus a relatively small depth effect – results for this specific scene. For $t_c = 48$ mm the parallax values are exactely twice as big as in the first case and, correspondingly, the perceived depth impression is also enlarged approximately by this factor.

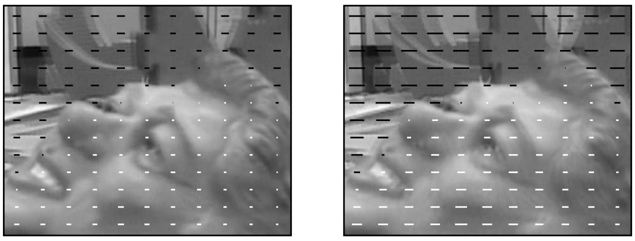
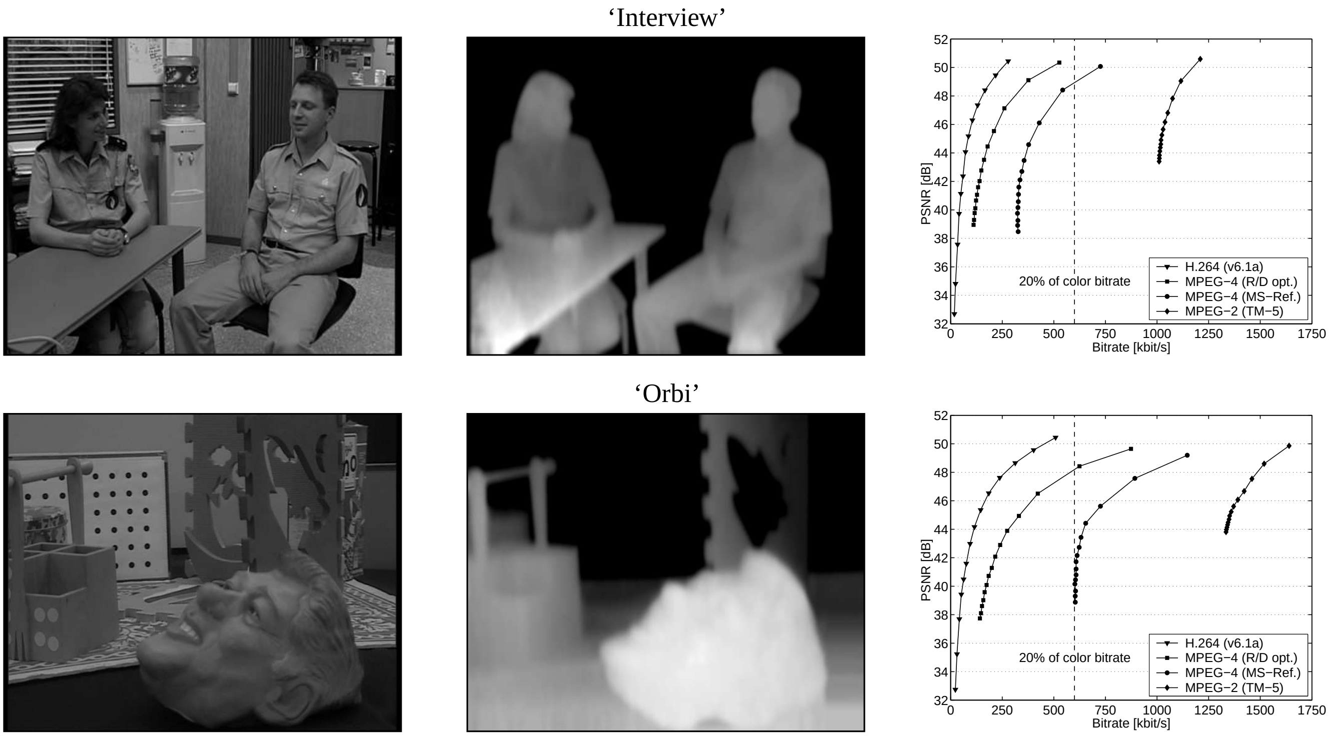


Figure 3: Magnified clippings from two "virtual" stereoscopic images synthesized with less, resp. more parallax, i.e. for a smaller, resp. a larger perceived depth effect.

3 Coding of 3D Imagery

For an in-depth stufy of the depth-image compression performance of the different MPEG technologies, three video coding standards were evaluated in a comparative coding


‘Interview’

‘Orbi’

Figure 4: Results of depth-image compression experiment for three different MPEG video coding standards.

experiment. The group of tested codecs consisted of: a) the MPEG-2 reference model codec (TM-5); b) the MPEG-4 Visual reference model codec from Microsoft® (MS-Ref.); c) a high-quality, rate-distortion (R/D) optimized MPEG-4 Visual codec developed at FhG/HHI (R/D opt.); d) the R/D optimized H.264 reference model codec a(v6.1a). The compression results for typical broadcast encoder settings, i.e. for a GOP (Group of Pictures) length equal to 12 with a GOP structure of *IBBPBBP...*, are shown in Fig. 4 for the two test sequences ‘Interview’ and ‘Orbi’.

The provided results show, first of all, that H.264 as well as MPEG-4 Visual seem to be very well suited for the coding of per-pixel depth information (with H.264 being even more efficient). The smoothness of the graylevel depth data 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, it can be followed from the rate-distortion (R/D) curves that the accompanying depth-images can be compressed to target rates significantly below 20% of this value. Just to give an example, H.264 compression of the ‘Interview’ sequence at 105 kbit/s still leads to a PSNR of about 46.29 dB. For the slightly more complex ‘Orbi’ scene, the same visual quality value can still be reached at a bitrate of approximately 184 kbit/s.

The quantitative results of the above-described coding experiment were also confirmed by means of subjective testing. In these trials, all in all 12 non-expert viewers were presented with “virtual” stereoscopic image material that was synthesized from more or less impaired encoded/decoded depth information [15]. Based on a subjective comparison with 3D imagery generated from the orginal, unimpaired depth data (‘best quality’ reference), the participants had to rate the stimuli in terms of: 1) perceived image impairments; 2) depth quality. Additionally, the subjects were asked to verbally describe perceived image distortions and 3D artifacts as well as their viewing experiences.

The rating results document that “virtual” stereoscopic images with an acceptable quality can be generated from very low bitrate depth information. In fact, it was found that even rather severe depth-image coding distortions such as visible blocking artifacts do not translate into equally strong perceptible impairments in the synthesized views.

At extremely low depth-image qualities it was, however, observed that synthesis distortions such as blocking artifacts, jagged object contours and depth layering (‘cardboarding’) appeared to be more pronounced in the foreground and tended to decrease in the background. This effect can be attributed to the *uniform* quantization of the employed per-pixel depth information; because of the non-linear ($\sim 1/Z$) depth-dependency of Eq. (1), the 3D warp does not lead to an equidistant spacing of the (discrete) parallax values in the “virtual” stereoscopic images. This again implicates that the resulting 3D rendering is reproduced on a 3D-TV display with a depth resolution that is much lower for near space regions than for objects that are further away. To counteract this unwanted nonlinear transformation charateristic, depth information should thus be quantized in a more suitable ('inverse') *nonuniform* fashion. This is described in detail in the following paragraph.

4 Optimal Depth Quantization

The optimal, nonuniform quantization can, for example, be derived from the theory of *plenoptic sampling*, which was first introduced by Chai *et al.* in 2000 [1]. In their influencial paper, the authors analyze the dependencies between camera distances, image- and depth resolution in the context of light-field rendering [9] for the case of an equidistant infinite camera array. They discover that "the spectral support of a 4D light-field is bounded by the minimum and maximum depths, irrespective of how complicated it might be because of depth variations in the scene". Virtually the same fundamental result – though expressed in much simpler, geometric terms – is provided by Feldmann *et al.* for the special case of a setup consisting of only two neighboring, parallel cameras [5]. Because their derivation is much easier to comprehend and to follow than the complex multidimensional spectral reasoning used in the plenoptic sampling theory, it shall also build the basis for the following discussions.

First of all, it must be noted that the *total* parallax range ΔP_{total} exhibited by a certain "virtual" stereoscopic image can be calculated from:

$$\Delta P_{total} = P\left(Z_{far}\right) - P\left(Z_{near}\right) \ , \tag{3}$$

where the two *bounding* parallax values $P\left(Z_{near}\right)$, resp. $P\left(Z_{far}\right)$, are due to scene parts at the near clipping plane Z_{near}, resp. at the far clipping plane Z_{far}, and are both calculated according to Eq. (1).

Now, assuming that the $N = 2^8$ discrete parallax values P provided by the "virtual" stereoscopic image indeed possess the desired equidistant spacing, the difference $\Delta P = |P_{\nu+1} - P_\nu|$ between parallax values $P_{\nu+1}$ and P_ν corresponding to neighboring depth layers $Z_{\nu+1}$ and Z_ν (with $\nu \in [0, \dots, N-2]$) can be calculated as:

$$\Delta P = \frac{\Delta P_{total}}{N-1} \ . \tag{4}$$

Then, any individual parallax value P_ν that corresponds to a certain discrete depth layer Z_ν (with $\nu \in [0, \dots, N-1]$) can be represented relative to the bounding parallax value $P\left(Z_{far}\right)$ as:

$$P_\nu = P\left(Z_{far}\right) - \nu \cdot \Delta P \ . \tag{5}$$

Solving this expression for Z_ν finally leads to the following conditional equation:

$$\frac{1}{Z_\nu} = \frac{1}{Z_{near}}\left(\frac{\nu}{N-1}\right) + \frac{1}{Z_{far}}\left(1 - \frac{\nu}{N-1}\right) \ , \tag{6}$$

which shows that the optimal, nonuniform depth quantization only depends on the *spatial extend* of the captured 3D scene, i.e. on the two clipping planes Z_{near} and Z_{far}, as well as on the total number of discrete depth layers N. This important relationship is also visualized in Fig. 5 in the following.[1]

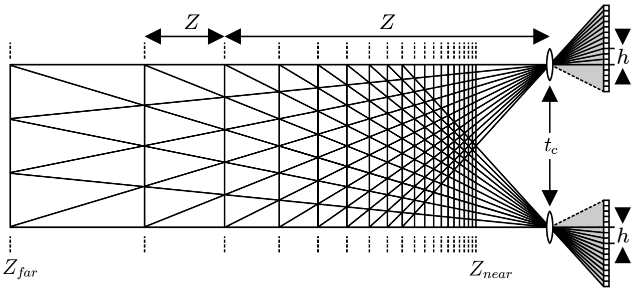


Figure 5: The optimal, nonuniform quantization of scene depth between the two clipping planes Z_{near} and Z_{far} results in the prefered equidistant spacing of parallax values (adapted from [17]).

It must be noted that in addition to the N *reconstruction levels* Z_ν defined by Eq. (6), the quantization process also requires a corresponding $N+1$ *decision levels* or *boundaries* $\mathcal{T}_\nu$. Therewith, any real-valued depth value Z can be assigned to a single, discrete depth layer Z_ν if $Z \in [\mathcal{T}_\nu, \mathcal{T}_{\nu+1})$, where the individual $\mathcal{T}_\nu$ are implicitely specified by:

$$-\frac{0.5}{N-1}\mathcal{Z}^{-1} \leq \frac{1}{Z} - \frac{1}{Z_\nu} < +\frac{0.5}{N-1}\mathcal{Z}^{-1} \ , \tag{7}$$

where $\mathcal{Z}^{-1} = (Z_{near}^{-1} - Z_{far}^{-1})$.

4.1 Influence on Coding & Synthesis

The influence of the two different depth quantization methods on the visual quality of synthesized 3D imagery was assessed in a comparative coding experiment. For that purpose, the floating-point depth information of the 'Cityhall' test sequence (see Fig. 6) was quantized between the near- and far clipping plane Z_{near}, resp. Z_{far}, both in the conventional uniform- and in the optimal, nonuniform fashion.

[1] It should be noted that this optimal, nonuniform quantization also results implicitely when the depth of a 3D scene is represented by means of *disparity vectors* between corresponding points in two images captured with a parallel stereo camera setup [17].

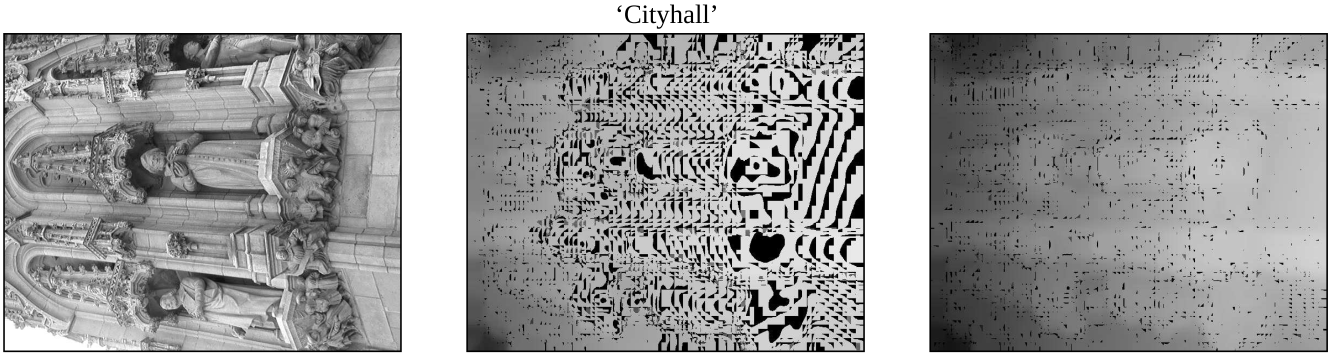


Figure 6: The optimal, nonuniform depth quantization (right) leads to a smaller amount of pixels with gross parallax errors (overlayed in black on the depth-images) than the conventional uniform spacing of depth values (center).

The two resulting 8-bit depth-image sequences were compressed at a range of different bitrates using the rate-distortion (R/D) optimized MPEG-4 Visual codec. "Virtual" stereoscopic images were then generated from the impaired depth information as well as from the monoscopic color video by means of the view synthesis algorithm ("virtual" stereo camera) developed in Sect. 2.

The quality of the synthesized 3D images might be expressed, for example, through the popular *mean squared error* (MSE), calculated between "virtual" views generated from original, resp. from impaired depth information [8]. The problem, however, lies in the fact that this value also depends to a large extend on the chosen resampling algorithm as well as on the texture of the 2D color video.

Therefore, the simpler *parallax error* is considered in the following. It is defined as:

$$\epsilon_P = P(p) - \hat{P}(p) \ , \tag{8}$$

where P, resp. $\hat{P}$, are the parallax values that correspond to the reference, resp. the encoded/decoded depth information and $p = (u, v)$ designates a pixel that lies within the respective image area $\mathcal{A}$.

Figure 7 shows that the optimal, nonuniform depth quantization actually leads to a smaller (root mean squared) parallax error:

$$\text{MSE}_{\mathcal{A}}(P, \hat{P}) = \frac{1}{|\mathcal{A}|} \sum_{p \in \mathcal{A}} |\epsilon_P|^2 \ , \tag{9}$$

than the conventional uniform quantization for the described 'Cityhall' sequence with its roughly equal amount of near- and far away scene parts.

Even more importantely, the optimal, nonuniform depth quantization also results in a smaller amount of pixels with *gross* parallax errors, i.e. with errors that exceed an experimentaly determined visibility threshold of $\beta = 0.8$ min of arc for the given viewing condition [7]. This is shown in Fig. 6 for a bitrate of about 430 kbit/s, where the 'heavily' impaired pixels are overlayed in black on the respective depth-images. (Here it is assumed that the stereoscopic imagery is observed from a viewing distance $D = 70$ cm on an 18" 3D display of width $W = 36$ cm.) Averaged over the whole sequence a total of 12.44% of the parallax values exceed β for the uniform quantization of depth, whereas only 4.27% of the image points produce gross parallax errors in case of the optimal, nonuniform depth quantization.

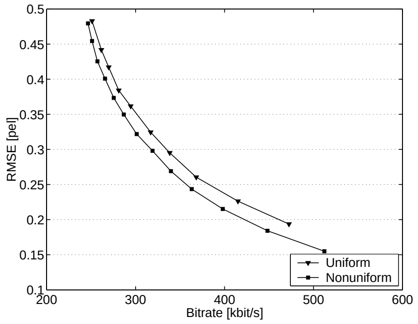


Figure 7: (Root mean squared) parallax errors for the two different types of depth quantization.

Acknowledgements

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 for their support and their input to this publication.

References

[1] J.-X. Chai, X. Tong, S. C. Chan, and H.-Y. Shum. Plenoptic Sampling. In *Proc. of ACM SIGGRAPH*,

pages 307–318, New Orleans, LA, USA, July 2000.

[2] 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, MA, USA, 2001.

[3] C. Fehn. Depth-Image-Based Rendering (DIBR), Compression and Transmission for a New Approach on 3D-TV. In *Proc. of SPIE Stereoscopic Displays and Applications XV*, pages 93–104, San Jose, CA, USA, January 2004.

[4] C. Fehn, K. Hopf, and B. Quante. Key Technologies for an Advanced 3D-TV System. To appear in *Proc. of SPIE Three-Dimensional TV, Video, and Display III*, Philadelphia, PA, USA, October 2004.

[5] I. Feldmann, O. Schreer, and P. Kauff. Nonlinear Depth Scaling for Immersive Video Applications. In *Proc. of 4th Int. Workshop on Image Analysis for Multimedia Interactive Services*, pages 433–438, London, UK, April 2003.

[6] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, UK, 2000.

[7] B. Kost and S. Pastoor. Visibility Thresholds for Disparity Quantization Errors in Stereoscopic Displays. In *Proc. of the Society of Information Display*, volume 32, number 2, pages 165–170, 1991.

[8] R. Krishnamurthy, B.-B. Chai, H. Tao, and S. Sethuraman. Compression and Transmission of Depth Maps for Image-Based Rendering. In *Proc. of Int. Conf. on Image Processing*, pages 828–831, Thessaloniki, Greece, October 2001.

[9] M. Levoy and P. Hanrahan. Light Field Rendering. In *Proc. of ACM SIGGRAPH*, pages 31–42, New Orleans, LA, USA, August 1996.

[10] L. Lipton. *Foundations of the Stereoscopic Cinema – A Study in Depth*. Van Nostrand Reinhold, New York, NY, USA, 1982.

[11] 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, April 1999.

[12] L. McMillan. *An Image-Based Approach to Three-Dimensional Computer Graphics*. PhD thesis, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, April 1997.

[13] 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 III*, pages 122–134, San Jose, CA, USA, February 1992.

[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(11):1335–1359, November 2000.

[15] B. Quante, C. Fehn, L. M. J. Meesters, P. J. H. Seuntiëns, and W. A. IJsselsteijn. Report on Perception of 3D Coding Artifacts. ATTEST Technical Report D8a, Eindhoven University of Technology and Fraunhofer Institute for Telecommunications, September 2003.

[16] A. Redert, M. Op de Beeck, C. Fehn, W. A. IJsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, I. Sexton, and P. Surman. ATTEST – Advanced Three-Dimensional Television System Technologies. In *Proc. of 1st Int. Symposium on 3D Data Processing, Visualization and Transmission*, pages 313–319, Padova, Italy, June 2002.

[17] D. Scharstein. *View Synthesis Using Stereo Vision*. PhD thesis, Cornell University, Ithaca, NY, USA, January 1997.

[18] H.-Y. Shum and S. B Kang. A Review of Image-Based Rendering Techniques. In *Proc. of Visual Communications and Image Processing*, pages 2–13, Perth, Australia, June 2000.

[19] A. Woods, T. Docherty, and R. Koch. Image Distortions in Stereoscopic Video Systems. In *Proc. of SPIE Stereoscopic Displays and Applications IV*, pages 36–48, San Jose, CA, USA, February 1993.

[20] G. Xu and Z. Zhang. *Epipolar Geometry in Stereo, Motion and Object Recognition*. Kluwer Academic Publishers, Dordrecht, The Netherlands, September 1996.