EFFICIENT DEPTH VIDEO CODING BASED ON VIEW SYNTHESIS DISTORTION ESTIMATION

Tae-Young Chung, Won-Dong Jang, and Chang-Su Kim

School of Electrical Engineering, Korea University, Seoul, Korea
E-mails: {tychung,wdjang}@mcl.korea.ac.kr, changsukim@korea.ac.kr

ABSTRACT

An efficient coding algorithm for depth map images and videos, based on view synthesis distortion estimation, is proposed in this work. We first analyze how a depth error is related to a disparity error and how the disparity vector error affects the energy spectral density of a synthesized color video in the frequency domain. Based on the analysis, we propose an estimation technique to predict the view synthesis distortion without requiring the actual synthesis of intermediate view frames. To encode the depth information efficiently, we employ a Lagrangian cost function to minimize the view synthesis distortion subject to the constraint on a transmission bit rate. In addition, we develop a quantization scheme for residual depth data, which adaptively assigns bits according to block complexities. Simulation results demonstrate that the proposed depth video coding algorithm provides significantly better R-D performance than conventional algorithms.

Index Terms— Multi-view plus depth, depth video coding, energy spectral density, and view synthesis distortion.

1. INTRODUCTION

With advances in multimedia technologies, especially the development of autostereoscopic 3D displays, multi-view video can support the rendering of a scene from various viewpoints and provide 3D perception with stereopsis [1] and motion parallax [2] cues. However, even though many efforts have been made for multi-view video compression [3], its transmission still requires a wide bandwidth and its bit-rate is almost linearly proportional to the number of views. To overcome this drawback without compromising stereopsis and motion parallax cues, multi-view video plus depth (MV+D) format was proposed [4]. An MV+D signal consists of a limited number of color videos, *e.g.* $2 \sim 3$ views, and their corresponding depth videos. Its bit-rate is lower than that of a multi-view video signal because of the limited number of views. An MV+D decoder can reconstruct arbitrary views between transmitted views based on the depth image based rendering (DIBR) [5]. DIBR converts a pixel in a view frame into 3D world coordinates with the respective depth information and camera parameters, and then re-projects the 3D coordinates into another pixel in a different view frame. Therefore, distortions in the depth information may lead to position errors in the view synthesis procedure, degrading the qualities of synthesized views severely.

This work was supported partly by the Global Frontier R&D Program on <Human-centered Interaction for Coexistence> funded by the National Research Foundation of Korea grant funded by the Korean Government(MEST) (NRF-M1AXA003-2011-0031648), and partly by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 2012-011031).

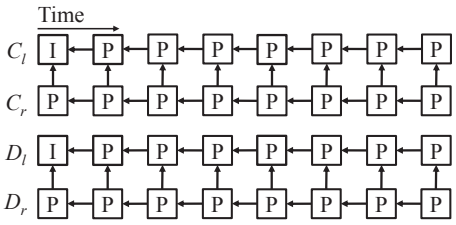

Fig. 1. A prediction structure for color videos plus their depth videos for two views. C_l and C_r denote left and right color videos, while D_l and D_r denote left and right depth videos.

Merkle *et al.* [6] showed that the distortions in depth image coding influence view synthesis performance strongly. Then, various attempts have been made to compress depth images efficiently by reducing view synthesis distortions instead of depth image distortions. Kim *et al.* [7] proposed a global linear function to estimate a view synthesis distortion from a depth distortion with two parameters, which describe the characteristics of a color video. They also proposed to skip the encoding of depth blocks when the corresponding color blocks are skipped. However, the performance improvement may be insignificant when the global function mismatches local characteristics of an input video signal. In [8], they proposed a local method to estimate a view synthesis distortion, assuming that local characteristics of a color video is similar to those of a synthesized video. Lee *et al.* [9] also proposed a skip mode for depth information coding, which determines skipped depth blocks based on the inter-view correlation between encoded color blocks. Their algorithm may be ineffective when the inter-view correlation is low, for example, due to illumination differences. Zhang *et al.* [10] proposed a view synthesis distortion estimation method, assuming that, in the frequency domain, the relation between depth distortion and view synthesis quality is similar to that between motion distortion and motion compensation quality [11]. But this method has limitations in applications, since it was developed to select the best intra mode only in the rate-distortion optimization in H.264/AVC.

In this work, we consider the color videos for two rectified views and their depth videos, as shown in Fig. 1. We first encode the color videos independently of the depth videos and then encode the depth videos. Therefore, the reconstructed color videos are available, when we encode the depth videos. We first derive the relationship between a depth error and a disparity error, and then analyze how the disparity error affects the energy spectral density of a synthesized color video. Based on the analysis, we propose a Lagrangian cost function to select the best encoding mode in the rate-distortion (R-D) sense. In addition, we develop an adaptive quantization scheme, which improves the performance of the depth video coding by assigning a larger amount of bits to more complicated blocks. Simulation re-

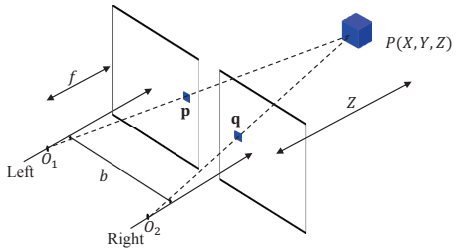

Fig. 2. Geometry relationship between an object in a 3D world coordinate system and its displacement in two 2D image coordinate systems.

sults demonstrate that the proposed algorithm provides significantly better R-D performance than the conventional algorithms [8, 10].

The rest of the paper is organized as follows. Section 2 analyzes the relation between a depth error, a disparity error, and a view synthesis distortion. Section 3 describes the proposed depth video coding algorithm. Section 4 evaluates the performance of the proposed algorithm. Finally, concluding remarks are given in Section 5.

2. ANALYSIS OF DEPTH ERROR AND VIEW SYNTHESIS DISTORTION

2.1. Depth Error vs. Disparity Error

In Fig. 2, a 3D point (X, Y, Z) in the world coordinate system is projected onto pixels $\mathbf{p}$ and $\mathbf{q}$ in left and right images, respectively. O_1 and O_2 are the optical centers of the left and right cameras. Since the two images are rectified, the y-position difference between $\mathbf{p}$ and $\mathbf{q}$ is zero. Also, when the focal length is f and the baseline distance between the cameras is b, the x-positions of $\mathbf{p}$ and $\mathbf{q}$ are given by $p_x = fX/Z$ and $q_x = f(X-b)/Z$. Therefore, the disparity vector $\mathbf{d_p}$ at pixel $\mathbf{p}$ is given by

$$\mathbf{d_p} = \mathbf{q} - \mathbf{p} = [-fb/Z, 0]^T. \tag{1}$$

In the MV+D format, we store the inverse of each object depth Z with 8-bit quantization. Let $Z_\mathbf{p}$ be the depth of pixel $\mathbf{p}$, and $z_\mathbf{p}$ be its inverse after the 8-bit quantization. The stored depth $z_\mathbf{p}$ has the range $[0, 255]$: 0 for the farthest point and 255 for the nearest point. The relationship between $Z_\mathbf{p}$ and $z_\mathbf{p}$ is given by

$$z_\mathbf{p} = 255 \times \frac{\frac{1}{Z_{\text{far}}} - \frac{1}{Z_\mathbf{p}}}{\frac{1}{Z_{\text{far}}} - \frac{1}{Z_{\text{near}}}}, \tag{2}$$

where Z_{far} and Z_{near} denote the farthest and the nearest depth in a scene. From Eqs. (1) and (2), the disparity vector $\mathbf{d_p}$ can be expressed as

$$\mathbf{d_p} = \left[fb \left(\frac{z_\mathbf{p}}{255} \left(\frac{1}{Z_{\text{near}}} - \frac{1}{Z_{\text{far}}} \right) + \frac{1}{Z_{\text{far}}} \right), 0 \right]^T. \tag{3}$$

Therefore, we can estimate the disparity error $\Delta\mathbf{d_p}$ from the depth error $\Delta z_\mathbf{p}$ by

$$\Delta\mathbf{d_p} = [\alpha \Delta z_\mathbf{p}, 0]^T \tag{4}$$

where $\alpha = fb\left(1/Z_{\text{near}} - 1/Z_{\text{far}}\right)/255$.

2.2. Disparity Error vs. View Synthesis Distortion

Suppose that block $\mathcal{B}_v$ from an arbitrary viewpoint v is synthesized from block $\mathcal{B}_l$ from the left viewpoint l by

$$\mathcal{B}_v(\mathbf{q}) = \mathcal{B}_l(\mathbf{p}), \tag{5}$$

where $\mathcal{B}_v(\mathbf{q})$ and $\mathcal{B}_l(\mathbf{p})$ denote the values of pixels $\mathbf{q}$ in $\mathcal{B}_v$ and $\mathbf{p}$ in $\mathcal{B}_l$, respectively, and $\mathbf{q} = \mathbf{p} + \mathbf{d_p}$. Then, assuming that the disparity errors of all pixels in $\mathcal{B}_l$ equal to $\Delta\mathbf{d}$, the distorted block $\tilde{\mathcal{B}}_v$ of $\mathcal{B}_v$ is given by

$$\tilde{\mathcal{B}}_v(\mathbf{q}) = \mathcal{B}_l(\mathbf{p} + \mathbf{d_p} + \Delta\mathbf{d}) = \mathcal{B}_v(\mathbf{q} + \Delta\mathbf{d}). \tag{6}$$

The Fourier transform decomposes a time-domain signal into its frequency components. By the shifting property of the Fourier transform, the translation of the time-domain signal corresponds to the phase shifting of the frequency domain signal. In addition, from the Parseval's relationship, the energy of the time-domain signal is identical with that of the frequency-domain signal. Therefore, the view synthesis distortion D_{view} of the reconstructed block $\tilde{\mathcal{B}}_v$, which is equivalent to the energy of the difference block $\mathcal{B}_v - \tilde{\mathcal{B}}_v$, can be computed in the frequency domain

$$D_{\text{view}} = \frac{1}{(2\pi)^2} \int\int S_l(\boldsymbol{\omega}) \left| 1 - e^{-j\boldsymbol{\omega}^T \Delta\mathbf{d}} \right|^2 d\omega_1 d\omega_2, \tag{7}$$

where $\boldsymbol{\omega} = [\omega_1, \omega_2]^T$ is the 2-D frequency vector and $S_l(\boldsymbol{\omega})$ denotes the energy spectral density of $\mathcal{B}_l$, which is the squared magnitude response of the Fourier transform of $\mathcal{B}_l$. The Taylor series expansion of $|1 - e^{-j\boldsymbol{\omega}^T \Delta\mathbf{d}}|^2$ yields

$$\begin{aligned} \left| 1 - e^{-j\boldsymbol{\omega}^T \Delta\mathbf{d}} \right|^2 &= 2 - 2\cos(\boldsymbol{\omega}^T \Delta\mathbf{d}) \\ &= \frac{2(\boldsymbol{\omega}^T \Delta\mathbf{d})^2}{2!} - \frac{2(\boldsymbol{\omega}^T \Delta\mathbf{d})^4}{4!} + \frac{6(\boldsymbol{\omega}^T \Delta\mathbf{d})^6}{6!} + \cdots \end{aligned} \tag{8}$$

As in [11], since the higher-order terms in Eq. (8) are insignificant for small $\boldsymbol{\omega}^T \Delta\mathbf{d}$, we can make the following approximation

$$D_{\text{view}} \simeq \frac{1}{(2\pi)^2} \int\int S_l(\boldsymbol{\omega}) \left(\boldsymbol{\omega}^T \Delta\mathbf{d} \right)^2 d\omega_1 d\omega_2. \tag{9}$$

Considering that the y-coordinate of the disparity error is zero, we further approximate the view synthesis distortion in Eq. (9) to

$$D_{\text{view}} \simeq \psi_x \left\| \Delta\mathbf{d} \right\|^2 \tag{10}$$

where

$$\psi_x = \frac{1}{(2\pi)^2} \int\int S_l(\boldsymbol{\omega}) \omega_1^2 d\omega_1 d\omega_2. \tag{11}$$

In this work, $\mathcal{B}_l$ is a 4×4 block. We hence replace the 2-D Fourier transform with the 4×4 discrete cosine transform (DCT) to compute the energy spectral density $S_l(\boldsymbol{\omega})$. The coefficients of the 4×4 DCT represent the signal strengths at discrete frequencies $(\pi u/4, \pi v/4)$, where $u = 0, 1, 2, 3$ and $v = 0, 1, 2, 3$. Therefore, by modifying Eq. (11) into a discrete form and setting ω_1 and ω_2 to $\pi u/4$ and $\pi v/4$, respectively, we can approximate ψ_x to

$$\psi_x = \frac{1}{64} \sum_{v=0}^{3} \sum_{u=0}^{3} H^2(u, v) u^2 \tag{12}$$

where $H(u, v)$ denotes the coefficient of the 4×4 DCT at (u, v).

Note that, in [10], Zhang *et al.* estimated the view synthesis distortion assuming that both x- and y-components of disparity errors equally contribute to the distortion. However, multi-view frames are generally rectified to reduce the two-dimensional correspondence problem to the one-dimensional problem. Thus, the y-components of disparity vectors are identically zero. Therefore, the assumption is

inappropriate, and the y-components are not related to the view synthesis distortion. Moreover, whereas Zhang *et al.*'s approximation uses the 2-D discrete Fourier transform to compute the energy spectral density, we employ the 4×4 DCT that is already implemented in the H.264/AVC codec.

3. PROPOSED DEPTH VIDEO CODING

To estimate the view synthesis distortion for a 16×16 block caused by depth errors, we first decompose the block into 16 sub-blocks of size 4×4. Let us suppose that all pixels in a sub-block have the same depth error. Then, from Eqs. (4), (10), and (12), we obtain the view synthesis distortion for the sub-block, given by

$$D_{\text{view}}^{4\times4} = \psi_x \times \frac{1}{16} \sum_{\mathbf{p}\in\mathcal{B}4} \alpha^2 |\Delta z_{\mathbf{p}}|^2. \tag{13}$$

In order to encode a 16×16 depth block, H.264/AVC employs the Lagrangian cost for the R-D optimization, which is given by

$$J_{\text{H.264}} = D_{\text{detph}}^{16\times16} + \lambda \cdot R^{16\times16} \tag{14}$$

where $D_{\text{detph}}^{16\times16}$ and $R^{16\times16}$ denote the depth distortion and the number of encoded bits for the 16×16 block. In this way, H.264/AVC considers the depth distortion instead of the view synthesis distortion, although the perceived video quality depends on the view synthesis distortion. Therefore, based on the formula in (13), we propose a more systematic Lagrangian cost J_{view}, which considers the view synthesis distortion $D_{\text{view}}^{16\times16}$ as follows.

$$J_{\text{view}} = D_{\text{view}}^{16\times16} + \lambda \cdot R^{16\times16} \tag{15}$$

where

$$D_{\text{view}}^{16\times16} = \sum_{i=0}^{15} D_{\text{view},i}^{4\times4}$$

and the subscript i is the index of a sub-block in the 16×16 block.

Furthermore, we develop an adaptive quantization scheme. Notice that residual depth data in a smooth region can be encoded with a large quantization parameter (QP) with little effect on the view synthesis distortion, whereas residual depth data in a complicated region should be encoded with a smaller QP. Let Q denote QP and $Q = Q_{\text{initial}} + \Delta Q$, where Q_{initial} is an initial QP and $\Delta Q = \{-2, -1, 0, 1, 2\}$. Then, we modify Eq. (15) to

$$J_{\text{view}} = D_{\text{view}}^{16\times16} + \lambda \cdot (R^{16\times16} + R_{\Delta Q}) \tag{16}$$

where $R_{\Delta Q}$ is the number of bits for the side information ΔQ. Based on this modified Lagrangian cost function, we determine the encoding mode, the motion vector, and the QP of a block adaptively, so that we minimize the distortion subject to the constraint on the limited bit budget.

4. SIMULATION RESULTS

We evaluate the performance of the proposed algorithm on four test sequences, which are listed in Table 1. These test sequences were recently released for the 3D video coding exploration experiments. The depth videos of these sequences are estimated by the depth estimation reference software [12], and then both color and depth videos are encoded by the JMVC 6.0 multi-view coding reference software [13] with the prediction structure in Fig. 1. Four quantization parameters were used: 27, 32, 37, and 42. An intermediate view is synthesized by the view synthesis reference software [12] using color videos and the corresponding depth videos. Each distorted intermediate view synthesized from reconstructed color and depth videos is compared with the lossless intermediate view synthesized from uncompressed color and depth videos. PSNR is used as the quality metric. We compare the proposed algorithm with three conventional depth video coding algorithms: H.264/AVC, Kim *et al.*'s algorithm [8], and Zhang *et al.*'s algorithm [10]. Note that we apply Zhang *et al.*'s algorithm to select the inter mode as well as the intra mode, even though it was originally developed for the intra mode decision only.

Table 1. Properties of four test sequences.

Sequence	Frame Size	Frame Rate (fps)	View number	
			Left	Right
Balloons	1024×768	30	3	5
BookArrival	1024×768	30	10	8
Lovebird1	1024×768	30	6	8
Pantomime	1280×960	30	39	41

Fig. 3 compares the rate-distortion curves of the proposed algorithm and the conventional algorithms on all test sequences, where horizontal and vertical axes denote the bit-rate for a depth video sequence and the average PSNR over all synthesized intermediate frames, respectively. Method I and Method II mean the proposed algorithm without and with the adaptive quantization scheme. We observe that Method I provides significantly better PSNR performance than the conventional algorithms. For example, on the "Balloons" sequence, Method I provides about 2.6 and 3.4 dB better PSNR's than Kim *et al.*'s algorithm and Zhang *et al.*'s algorithm, respectively. Although only x-components of disparity vectors are corrupted by depth distortions, Zhang *et al.*'s algorithm considers the energy spectral density due to the errors in both x and y directions. On the contrary, the proposed algorithm considers the errors in x direction only. Therefore, the proposed algorithm estimates view synthesis distortions more accurately and outperforms Zhang *et al.*'s algorithm in terms of the R-D performance. Moreover, Method II improves the performance even further by assigning a larger amount of bits to blocks with higher energy spectral densities.

Table 2 provides the Bjontegarrd evaluation results to measure the average bit-rate reduction or PSNR increase, when the H.264/AVC standard is used as the benchmark. We see that Method I reduces the average bit-rate by about 74.9% or increases the average PSNR by about 4.89 dB. Furthermore, Method II reduces the average bit-rate by 81.0% or increases the average PSNR by 6.48 dB. On the other hand, Kim *et al.*'s algorithm and Zhang *et al.*'s algorithm reduce the average bit-rate by 54.9% and 63.7% or increases the average PSNR by 3.29 dB and 3.03 dB, respectively. These simulation results indicate that the proposed algorithm outperforms the conventional algorithms by a large margin.

5. CONCLUSIONS

We proposed the R-D optimized depth video coding algorithm and the adaptive quantization scheme, based on the view synthesis distortion estimation. We first derived the relationship between a depth error and a disparity error, and analyzed how the disparity error affects a view synthesis distortion in the frequency domain. Through the analysis, we showed that the view synthesis distortion can be estimated from the energy spectral density of a color video signal. Then, we proposed the Lagrangian cost function to minimize the view syn-

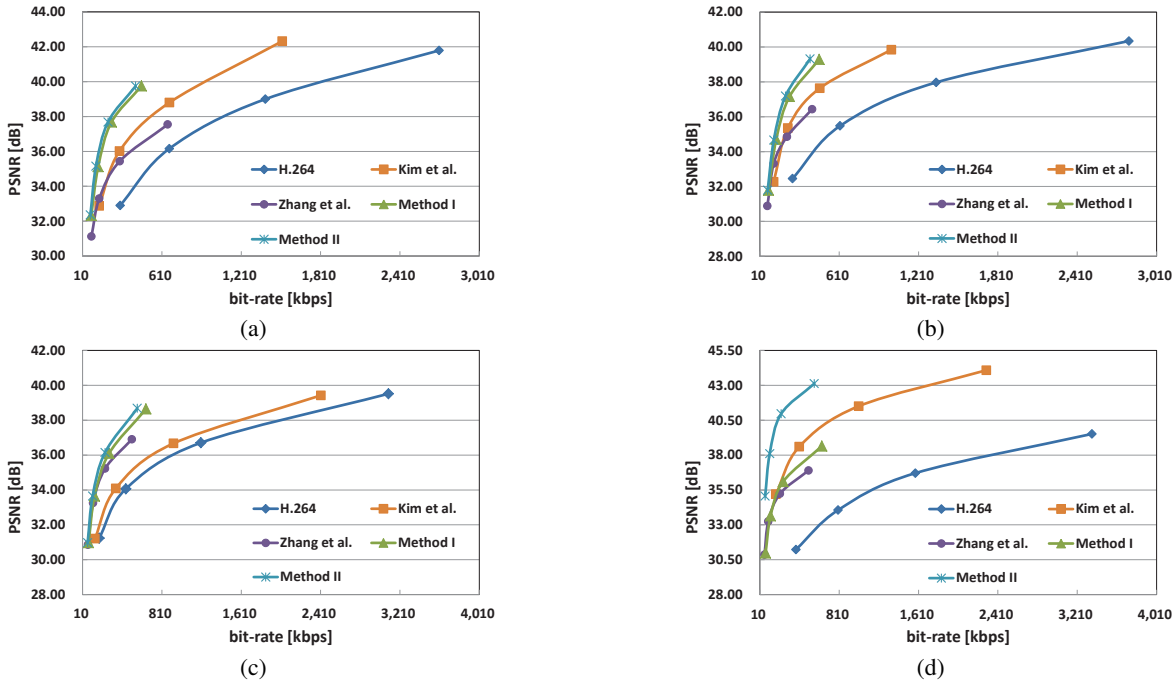

Fig. 3. Comparison of the R-D curves of the proposed algorithm with H.264/AVC, Kim *et al.*'s algorithm [8], and Zhang *et al.*'s algorithm [10]: (a) "Balloons," (b) "BookArrival," (c) "Lovebird1," and (d) "Pantomime" sequences.

Table 2. The average bit rate and PSNR differences from the benchmark, which is the H.264/AVC standard. A negative value means a bit rate decrement in comparison with the benchmark, whereas a positive value means a PSNR increment.

Sequences	Kim *et al.* [8]		Zhang *et al.* [10]		Method I		Method II	
	ΔBits [%]	ΔPSNR [dB]	ΔBits [%]	ΔPSNR [dB]	ΔBits [%]	ΔPSNR [dB]	ΔBits [%]	ΔPSNR [dB
Balloons	-52.14	2.78	-48.17	1.96	-74.11	5.40	-78.12	5.72
BookArrival	-60.76	3.07	-58.85	2.52	-74.48	4.98	-77.87	5.50
Lovebird1	-22.45	0.72	-64.46	2.85	-68.06	3.44	-72.98	3.94
Pantomime	-84.42	6.58	-72.98	3.94	-81.75	5.74	-94.85	10.79

thesis distortion subject to the constraint on a bit-rate. Moreover, we developed the adaptive quantization scheme to improve the R-D performance further by assigning a larger amount of bits to more complicated blocks. Simulation results confirmed that the proposed algorithm provides much better R-D performance than the conventional algorithms [8, 10].

6. REFERENCES

[1] I. P. Howard and B. J. Rogers, *Binocular Vision and Stereopsis*, Oxford University Press, 1995.

[2] I. P. Howard and B. J. Rogers, "Seeing in depth: Depth perception (vol. 2)," *Toronto: I. Porteus*, 2002.

[3] M. Flierl and B. Girod, "Multiview video compression," *IEEE Signal Process. Mag.*, vol. 24, no. 6, pp. 66–76, Nov. 2007.

[4] ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, "Multi-view Video plus Depth (MVD) Format for Advanced 3D Video System," Doc. JVT-W100, Apr. 2007.

[5] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," in *Proc. SPIE Steteoscopic Displays and Virtual Reality Systems XI*, 2004, pp. 93–104.

[6] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Proc. IEEE ICIP*, Sept. 2007, pp. 201–204.

[7] W. S. Kim, A. Ortega, P. L. Lai, D. Tian, and C. Gomila, "Depth map distortion analysis for view rendering and depth coding," in *Proc. IEEE ICIP*, Nov. 2009, pp. 721–724.

[8] W. S. Kim, A. Ortega, P. L. Lai, D. Tian, and C. Gomila, "Depth map coding with distortion estimation of rendered view," in *Proc. SPIE VCIP*, Jan. 2010, p. 75430B.

[9] J. Y. Lee, H. Wey, and D. S. Park, "A novel approach for efficient multi-view depth map coding," in *Proc. Picture Coding Symposium*, Dec. 2010, pp. 302–305.

[10] Q. Zhang, P. An, Y. Zhang, and Z. Zhang, "Efficient rendering distortion estimation for depth map compression," in *Proc. IEEE ICIP*, Sept. 2011, pp. 1129–1132.

[11] A. Secker and D. Taubman, "Highly scalable video compression with scalable motion coding," *IEEE Trans. on Image Process.*, vol. 13, no. 8, pp. 1029–1041, Aug. 2004.

[12] ISO/IEC JTC1/SC29/WG11, "Reference Software for Depth Estimation and View Ssynthesis," Doc. N15377, Apr. 2008.

[13] ISO/IEC JTC1/SC29/WG11, "Study Test of ISO/IEC 14496-5:2001/FPDAM 15 Reference Software for Multiview Video Coding," Doc. N10704, Jul. 2009.