Joint Source-Channel Coding and Unequal Error Protection for Video Plus Depth

Arash Vosoughi, *Student Member, IEEE*, Pamela C. Cosman, *Fellow, IEEE*, and Laurence B. Milstein, *Fellow, IEEE*

***Abstract*—We consider the joint source-channel coding (JSCC) problem of 3-D stereo video transmission in video plus depth format over noisy channels. Full resolution and downsampled depth maps are considered. The proposed JSCC scheme yields the optimum color and depth quantization parameters as well as the optimum forward error correction code rates used for unequal error protection (UEP) at the packet level. Different coding scenarios are compared and the UEP gain over equal error protection is quantified for flat Rayleigh fading channels.**

***Index Terms*—3-D video, joint source-channel coding, unequal error protection, video plus depth.**

I. Introduction

WE ARE interested in the delivery of three-dimensional (3-D) video over mobile devices [1]. The quality of a received 3-D video is affected by both the source coding accuracy and the amount of redundancy introduced by forward error correction (FEC) to protect the compressed 3-D video transmitted over a channel. Therefore, for a fixed bitrate, designing a clever method to divide the bits between the source and FEC is crucial in order to maximize the quality at the receiver. This problem is called joint source-channel coding (JSCC) and it is a well-studied area for 2D video.

Video plus depth (V+D) is an efficient representation of 3-D video, where a stereo pair is rendered at the decoder from a color video signal and a per-pixel depth map [1], [2]. In [3], two different protection levels are considered for V+D, and the authors concluded that color should be protected more strongly than depth. Following this conclusion, a UEP method is proposed in [4] for V+D data over WiMAX communication channels based on unequal power allocation. In [5], it is concluded that depth can be compressed more compared to color, and downsampling the depth by a factor of two is recommended to increase coding efficiency, although the effect of a channel is not investigated.

In this paper, the JSCC problem is solved for V+D data transmitted over noisy channels. We consider both downsampled and full resolution depth scenarios. Both the color and depth are encoded by an H.264/AVC encoder [6] and then protected by FEC using UEP such that each individual packet is protected according to its importance. The importance of packets is based on the structural similarity (SSIM) index [7]. The JSCC yields the optimum color and depth quantization parameters as well as the UEP code rates that jointly maximize the quality at the receiver. Turbo codes [8] are used for FEC, and simulation results are given for flat Rayleigh fading channels. The performances of different scenarios are compared, and UEP performance is compared to EEP.

Manuscript received June 20, 2014; accepted July 25, 2014. Date of publication August 12, 2014; date of current version August 20, 2014. This work was supported by the Intel/Cisco Video Aware Wireless Networks (VAWN) program and by the National Science Foundation under Grant CCF-1160832. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Oscar C. Au.

The authors are with the University of California, San Diego, Department of Electrical and Computer Engineering, La Jolla, CA 92093-0407 USA (e-mail: arvosoughi@ucsd.edu; pcosman@ucsd.edu; lmilstein@ucsd.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2014.2346739

II. JSCC Problem Formulation

The system block diagram is shown in Fig. 1. The color and depth are both compressed by an H.264/AVC encoder, protected by FEC (using UEP), and then transmitted over a channel. At the receiver, the erroneous packets are detected and discarded after channel decoding, and the color (left view) and depth bitstreams are decoded by an H.264/AVC decoder, where error concealment (EC) is done for the discarded (lost) packets. The right view is then synthesized using the decoded color and depth. The depth map may be downsampled by a factor of M before the compression, and thus should be upsampled by a factor of M before the view synthesis. We consider full resolution and downsampled depth by factors of 2 and 4, which are represented by $\downarrow \mathrm{No}, \downarrow 2$, and $\downarrow 4$, respectively. In this work, we use SSIM since perceived quality is better correlated to SSIM than to PSNR [9]. The SSIM between two GOPs (group of pictures) x and y is calculated as the average of SSIMs between the corresponding frames of x and y, and is denoted by $\mathrm{SSIM}(x, y)$. It varies between -1 and 1, where larger values correspond to lower distortion. We first derive a measure of end-to-end distortion for the left (color) view based on SSIM. This measure should incorporate both the effects of the color source distortion and the color channel distortion. Each packet of the color GOP is assigned a score. The score depends on whether the packet is or is not lost. If the ith packet of a color GOP is lost, the score assigned to that packet is

$$d^L_{i,C}(q_C) = \mathrm{SSIM}(f^L, \tilde{f}^L_{i,C}(q_C)), \tag{1}$$

where q_C is the color quantization parameter, f^L denotes the original uncompressed left view GOP, and $\tilde{f}^L_{i,C}(q_C)$ represents the left view decoded GOP with error concealment as if only the ith color packet is lost. We note that $d^L_{i,C}(q_C)$ reflects the quality throughout the left view GOP (including the effect of error propagation) due to losing the ith color packet; larger values of $d^L_{i,C}(q_C)$ correspond to lower distortion generated due to loss of the ith packet. If the ith color packet is not lost, the score assigned to that packet is

$$d^L_s(q_C) = \mathrm{SSIM}(f^L, \hat{f}^L(q_C)), \tag{2}$$

where $\hat{f}^L(q_C)$ denotes the left view error-free decoded GOP. Since each packet has two different scores (given in (1) and (2)), the score is a random variable. Let $\mathbf{D}_i$ be the random variable

representing the score assigned to the ith packet. Thus, $\{\mathbf{D}_i = d^L_{i,C}(q_C)\}$ if the packet is lost, and $\{\mathbf{D}_i = d^L_s(q_C)\}$ if the packet is not lost. For a particular q_C, $d^L_{i,C}(q_C)$ and $d^L_s(q_C)$ can be computed offline at the encoder for $1 \leq i \leq N_C$, where N_C is the number of packets in a color GOP. We take the expected value of the average of the scores of all the color packets as the quality of the left view:

$$E^L = E\left\{\frac{1}{N_C}\sum_{i=1}^{N_C}\mathbf{D}_i\right\}. \tag{3}$$

Let $p_{i,C}\left(s_{i,C}\left(q_C\right), r_{i,C}, \Theta\right)$ be the probability of losing the ith color packet, where $p_{i,C}$ depends on the source packet size in bits, $s_{i,C}(q_C)$, the code rate allocated to that packet, $r_{i,C}$, and the channel characteristics Θ. For a flat Rayleigh fading channel, $\Theta = (\mathrm{SNR}, T_c)$, where T_c is the coherence time defined in Section III. Following (3), we have

$$E^L = \frac{1}{N_C}\sum_{i=1}^{N_C}\left(d^L_s\left(q_C\right) + p_{i,C}\left(s_{i,C}\left(q_C\right), r_{i,C}, \Theta\right)\cdot \left(d^L_{i,C}(q_C) - d^L_s\left(q_C\right)\right)\right). \tag{4}$$

For the synthesized right view, scores are defined for both color and depth packets, since both contribute to the quality of the synthesized right view. If the ith color packet is lost, the assigned score is

$$d^R_{i,C}(q_C, q_D) = \mathrm{SSIM}(f^R, \tilde{f}^R_{i,C}(q_C, q_D)), \tag{5}$$

and if the ith depth packet is lost, the assigned score is

$$d^R_{i,D}(q_C, q_D) = \mathrm{SSIM}(f^R, \tilde{f}^R_{i,D}(q_C, q_D)). \tag{6}$$

In (5) and (6), q_D is the depth quantization parameter, f^R is the GOP synthesized from the original left view ([10], [11]), $\tilde{f}^R_{i,C}(q_C, q_D)$ denotes the right view GOP after decoding, EC, and view synthesis as if only the ith packet is lost from the color, and $\tilde{f}^R_{i,D}(q_C, q_D)$ denotes the right view GOP after decoding, EC, and view synthesis as if only the ith packet is lost from the depth. If the ith packet of the color or the depth is not lost, the score assigned to that packet is

$$d^R_s(q_C, q_D) = \mathrm{SSIM}(f^R, \hat{f}^R(q_C, q_D)), \tag{7}$$

where $\hat{f}^R(q_C, q_D)$ denotes the error-free decoded synthesized right view GOP. Similar to (3), we consider the expected value of the average of scores of color and depth packets as the quality of the synthesized right view:

$$E^R = \frac{1}{N_C+N_D}\left(\sum_{i=1}^{N_C}\left(d^R_s\left(q_C, q_D\right)+p_{i,C}\left(s_{i,C}\left(q_C\right), r_{i,C}, \Theta\right)\cdot \left(d^R_{i,C}\left(q_C, q_D\right) - d^R_s\left(q_C, q_D\right)\right)\right) + \sum_{i=1}^{N_D}\left(d^R_s\left(q_C, q_D\right) + p_{i,D}\left(s_{i,D}\left(q_D\right), r_{i,D}, \Theta\right)\left(d^R_{i,D}\left(q_C, q_D\right) - d^R_s\left(q_C, q_D\right)\right)\right), \tag{8}$$

where N_D is the number of packets of a depth GOP, $s_{i,D}(q_D)$ is the size of the ith depth source packet in bits, and $r_{i,D}$ and $p_{i,D}$ are, respectively, the code rate allocated to that packet and the probability of losing that packet. We define the objective function of the JSCC problem as $\frac{E^L+E^R}{2}$. An interpretation of this objective function is as follows: Let us consider the ith and the jth packets of the color, where $i \neq j$ and $1 \leq i, j \leq N_C$. The contribution of the ith packet and the jth packet to the E^L term of the objective function is equal to $f_i = \frac{d^L_s(q_C)+p_{i,C}\times\left(d^L_{i,C}(q_C)-d^L_s(q_C)\right)}{N_C}$ and $f_j = \frac{d^L_s(q_C)+p_{j,C}\times\left(d^L_{j,C}(q_C)-d^L_s(q_C)\right)}{N_C}$, respectively. We note that $d^L_{k,C}(q_C) - d^L_s(q_C) \leq 0$ for $1 \leq k \leq N_C$. Thus, if $p_{i,C} = p_{j,C}$ and $d^L_{i,C}(q_C) > d^L_{j,C}(q_C)$, or, if $p_{i,C} < p_{j,C}$ and $d^L_{i,C}(q_C) = d^L_{j,C}(q_C)$, then $f_i > f_j$. This means that a packet with a lower distortion value (larger score) or a smaller loss probability has a larger contribution to the objective function, that is to be maximized. Further, note that if $d^L_{i,C}(q_C) = d^L_s(q_C)$, then $f_i = \frac{d^L_s(q_C)}{N_C}$, meaning that the contribution of a packet with no channel distortion due to error concealment is equal to the source distortion averaged over all the packets. The interpretation given above is for the E^L term of the objective function; a similar interpretation can be made for the E^R term.

The total number of bits, which is the sum of the number of source bits and FEC bits, is equal to $\sum_{i=1}^{N_C}\frac{s_{i,C}(q_C)}{r_{i,C}} + \sum_{i=1}^{N_D}\frac{s_{i,D}(q_D)}{r_{i,D}}$. Let $\mathcal{R}$ be the set of available code rates, and $\mathcal{Q}_C$ and $\mathcal{Q}_D$ represent the sets of quantization parameters used to encode the color and depth, respectively. Let $\mathbf{R_C} \triangleq (r_{1,C}, r_{2,C}, \ldots, r_{N_C,C})$, and $\mathbf{R_D} \triangleq (r_{1,D}, r_{2,D}, \ldots, r_{N_D,D})$. To maximize the quality of the received 3-D video, we maximize the objective function

$$\max_{\substack{(q_C, q_D) \in \mathcal{Q}_C \times \mathcal{Q}_D \\ \mathbf{R_C} \in \mathcal{R}^{N_C}, \mathbf{R_D} \in \mathcal{R}^{N_D}}} \frac{E^L + E^R}{2}, \tag{9}$$

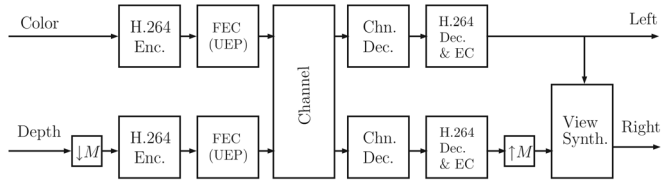

Fig. 1. System block diagram.

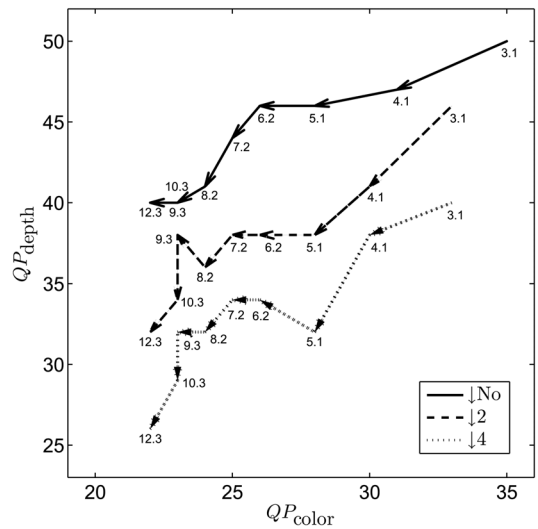

Fig. 2. Trajectories of the optimum QPs for ↓ No, ↓ 2, and ↓ 4 scenarios for a flat Rayleigh fading channel with SNR=8 dB and $T_c = 4000$. Numbers next to the trajectories denote the bitrate constraints in Mb/sec.

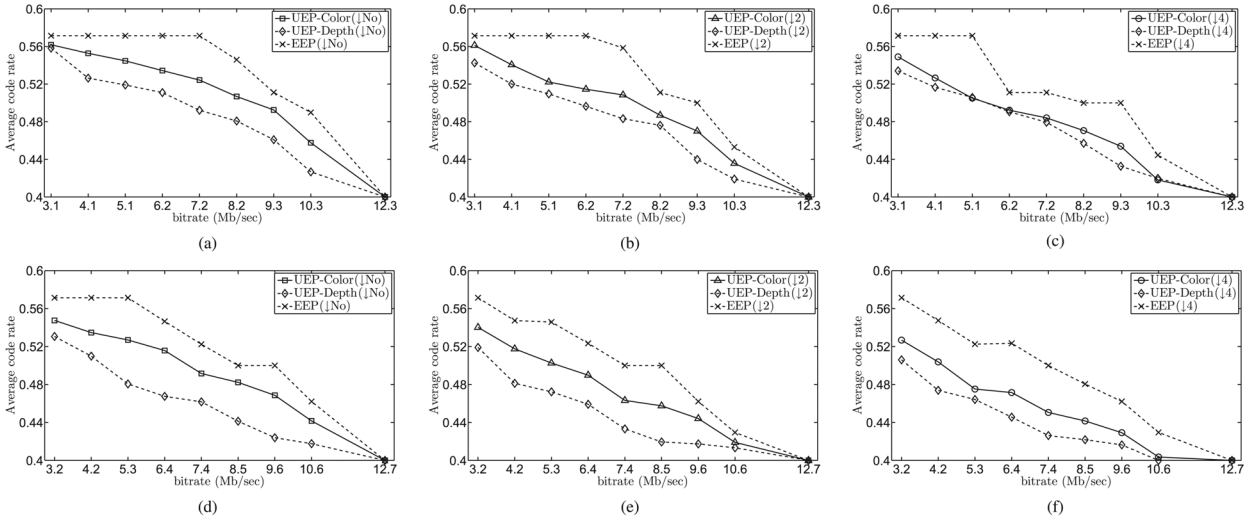

Fig. 3. Average color and depth code rates for a flat Rayleigh fading channel with $\mathrm{SNR} = 8$ dB. (a)-(c) 'Balloons', and (d)-(f) 'Poznanstreet'.

subject to the bit constraint

$$\sum_{i=1}^{N_C} \frac{s_{i,C}(q_C)}{r_{i,C}} + \sum_{i=1}^{N_D} \frac{s_{i,D}(q_D)}{r_{i,D}} \leq B, \tag{10}$$

where B is the bit budget. It is assumed that the channel is known at the transmitter, meaning that an accurate estimate of Θ is available at the transmitter side. The optimization problem introduced in (9) and (10) is a discrete optimization problem that is solved using the branch and bound method [12].

III. Simulation Results and Discussion

We present simulation results for flat Rayleigh fading channels with binary phase-shift keying (BPSK) modulation/demodulation. The coherence time of a fading channel, T_c, represents the number of symbols affected by the same fade level, and using a block-fading model, each fade is considered to be independent of the others. An interleaver mitigates the effect of error bursts, and we use a block interleaver with depth 500 and width 100. We use UMTS turbo codes for FEC [13]. The available code rates we considered are $\{\frac{8}{9}, \frac{4}{5}, \frac{2}{3}, \frac{4}{7}, \frac{1}{2}, \frac{4}{9}, \frac{2}{5}, \frac{4}{11}, \frac{1}{3}\}$, obtained by puncturing a mother code of rate $\frac{1}{3}$. An iterative soft-input/soft-output (SISO) decoding algorithm is used for turbo decoding. We used H.264/AVC reference software (JM version 15.1), with motion compensated error concealment (MCEC). Each row of macroblocks is encoded as a packet, the GOP structure is IPPP, and the GOP size is 10 frames.

Fig. 2 shows the trajectories of the optimum QPs obtained for a GOP of 'Balloons' video sequence (1024×768) as the bitrate constraint increases, where $\mathrm{SNR} = 8$ dB and $T_c = 4000$. For the $\downarrow$ No scenario, when the bitrate constraint is 3.1 Mbps, $QP_{\mathrm{depth}} = 50$ and $QP_{\mathrm{depth}} = 35$. When the bitrate constraint increases to 12.3 Mbps, QP_{depth} goes to 40 and QP_{color} goes to 22. This shows that over a range of rates, the depth can be significantly compressed compared to the color. When the depth is downsampled, QP_{depth} is still larger than QP_{color}, but the gap is smaller than for the $\downarrow$ No scenario, showing that when depth has lost spatial resolution, the optimization does not penalize it so much on compression.

Fig. 3 shows the color and depth average code rates versus the bitrate constraint for video sequences 'Balloons' and 'Poznanstreet' (1920×1088). The average code rates decrease when the bitrate constraint increases, meaning that more protection is provided for both by increasing the bitrate. Considering the $\downarrow$ No scenario, we see that although the depth is significantly compressed compared to the color (see Fig. 2), it is protected more since the depth average code rate is lower than that of the color. In [3], the authors concluded that color should be protected more than depth. That conclusion was made for the symmetric coding case, where $QP_{\mathrm{color}} = QP_{\mathrm{depth}} = 30$. We also solved the JSCC problem with the additional symmetric coding constraint, i.e., we set $q_C = q_D$ in (9) and (10), and our results showed that, indeed, the color should be protected more than the depth, in agreement with [3]. In other words, we are in agreement with the results of [3] for the special case of identical quantization parameters, but the general case of unequal quantization parameters yields the result that depth should be compressed more severely than color, but that then depth should be protected more. Results for $\downarrow 2$ and $\downarrow 4$ scenarios also indicate that the JSCC tends to protect the depth slightly more than the color.

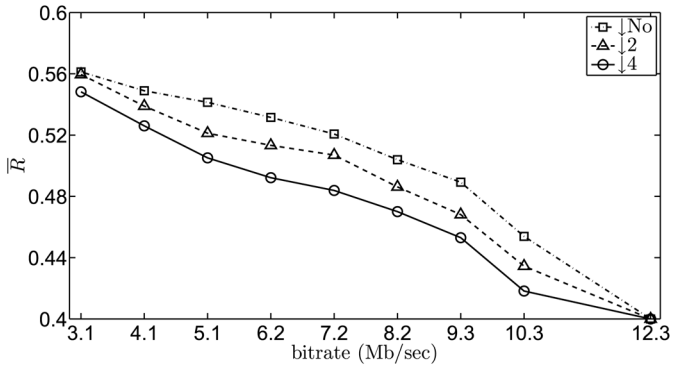

Fig. 4. $\overline{R}$ for $\downarrow$ No, $\downarrow 2$, and $\downarrow 4$ scenarios for a flat Rayleigh fading channel with $\mathrm{SNR} = 8$ dB and $T_c = 4000$.

We now compare the scenarios $\downarrow$ No, $\downarrow 2$, and $\downarrow 4$ in terms of FEC protection. We compute the average code rate $\overline{R}$:

$$\overline{R} = \frac{\#\text{color source bits} + \#\text{depth source bits}}{\#\text{color source} + \text{FEC bits} + \#\text{depth source} + \text{FEC bits}}.$$

Fig. 4 shows $\overline{R}$ versus the bitrate constraint for 'Balloons'. For a particular bitrate, $\overline{R}$ decreases when the downsampling factor increases, meaning that for the same bitrate constraint, a stronger protection is needed for a larger downsampling factor.

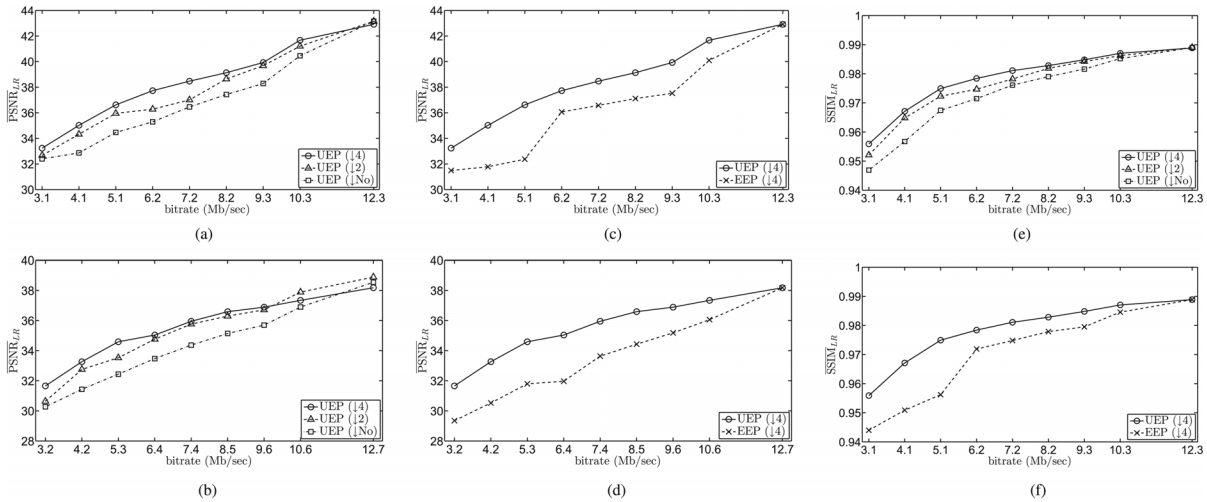

Fig. 5. (a)-(b) $\overline{\mathrm{PSNR}}_{LR}$ for scenarios $\downarrow \mathrm{No}$, $\downarrow 2$, and $\downarrow 4$, (c)-(d) $\overline{\mathrm{PSNR}}_{LR}$ of UEP and EEP for scenario $\downarrow 4$, (e) $\overline{\mathrm{SSIM}}_{LR}$ for scenarios $\downarrow \mathrm{No}$, $\downarrow 2$ and $\downarrow 4$, and (f) $\overline{\mathrm{SSIM}}_{LR}$ of UEP and EEP for scenario $\downarrow 4$.

Different scenarios are also compared for channel realizations using the PSNR and SSIM metrics. Following [10] and [11], in computing the full-reference metrics PSNR and SSIM, the reference of the right view is obtained by view synthesis from the original uncompressed left view. For each channel realization, the left and right view SSIMs are averaged and then the average is taken over all the channel realizations, which is denoted by $\overline{\mathrm{SSIM}}_{LR}$. The average PSNR for each channel realization is calculated by $10\log_{10}(\frac{255^2}{(\mathrm{MSE}_L+\mathrm{MSE}_R)/2})$, where MSE_L and MSE_R are the mean-squared errors obtained for the left and right views, respectively. The average is then taken over all the channel realizations, which is denoted by $\overline{\mathrm{PSNR}}_{LR}$. Figs. 5(a) and (b) show $\overline{\mathrm{PSNR}}_{LR}$ for 200 channel realizations for 'Balloons' and 'Poznanstreet', respectively, where SNR=8 dB and $T_c = 4000$. Results for $\overline{\mathrm{SSIM}}_{LR}$ are given in Fig. 5(e) for 'Balloons'. We see that the $\downarrow 4$ scenario outperforms the other scenarios except for high bitrates, for which the $\downarrow 2$ scenario slightly outperforms the others. Similar results (not shown) were obtained for video sequence 'Mobile'.

Lastly, we compare the performance of UEP to that of EEP. Results are given for scenario $\downarrow 4$, which was the best for most of the bitrates and channel conditions considered. Figs. 5(c) and (d) show $\overline{\mathrm{PSNR}}_{LR}$ for 'Balloons' and 'Poznanstreet', respectively. We see that UEP outperforms EEP by up to 4.3 dB and 3.1 dB for 'Balloons' and 'Poznanstreet', respectively. Fig. 5(f) shows $\overline{\mathrm{SSIM}}_{LR}$ for 'Balloons'. UEP outperforms EEP in terms of SSIM as well.

IV. Conclusions

JSCC was studied for video plus depth. Full resolution and downsampled depth by factors of two and four were considered. Results show that the depth can be significantly compressed compared to the color (especially for $\downarrow \mathrm{No}$ and $\downarrow 2$), although it needs to be protected more by FEC. We showed that when depth is downsampled, it should be less compressed and more protected to maximize the quality. In contrast to prior work which only considered equal quantization parameters and found that color should be more protected than depth, we found that depth should be compressed more severely than color and then protected more. We also showed that the downsampled depth by a factor of four outperforms the other scenarios except for high bitrates. The UEP approach proposed here was shown to yield up to 4.3 dB gain in PSNR compared to EEP for flat Rayleigh fading channels.

References

[1] A. Gotchev, G. Akar, T. Capin, D. Strohmeier, and A. Boev, "Three-dimensional media for mobile devices," *Proc. IEEE*, vol. 99, no. 4, pp. 708–741, 2011.

[2] A. Smolic, K. Mueller, P. Merkle, P. Kauff, and T. Wiegand, "An overview of available and emerging 3D video formats and depth enhanced stereo as efficient generic solution," in *Picture Coding Symp., 2009 PCS 2009*, 2009, pp. 1–4.

[3] C. T. E. R. Hewage, S. Worrall, S. Dogan, H. Kodikaraarachchi, and A. Kondoz, "Stereoscopic TV over IP," in *2007 IETCVMP. 4th Eur. Conf. Visual Media Production*, Nov. 2007, pp. 1–7.

[4] C. T. E. R. Hewage, Z. Ahmad, S. Worrall, S. Dogan, W. A. C. Fernando, and A. Kondoz, "Unequal Error Protection for backward compatible 3-D video transmission over WiMAX," in *IEEE Int. Symp. Circuits and Systems 2009 ISCAS 2009*, May 2009, pp. 125–128.

[5] K. Klimaszewski, K. Wegner, and M. Domanski, "Influence of views and depth compression onto quality of synthesized view," 2009, ISO/IEC JTC1/SC29/WG11 MPEG2009/M16758. London, U.K..

[6] T. Wiegand, G. Sullivan, G. Bjontegaard, 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.

[7] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[8] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes(1)," in *IEEE Int. Conf. Commun.*, May 1993, vol. 2, pp. 1064–1070.

[9] P. Hanhart and T. Ebrahimi, "Quality assessment of a stereo pair formed from decoded and synthesized views using objective metrics," in *3DTV-Conf.: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2012*, Oct. 2012, pp. 1–4.

[10] A. Tikanmaki, A. Gotchev, A. Smolic, and K. Miller, "Quality assessment of 3D video in rate allocation experiments," in *IEEE Int. Symp. Consumer Electronics*, Apr. 2008, pp. 1–4.

[11] G. Tech, A. Smolic, H. Brust, P. Merkle, K. Dix, Y. Wang, K. Müller, and T. Wiegand, "Optimization and comparison of coding algorithms for mobile 3DTV," in *3DTV Conf.: The True Vision-Capture, Transmission and Display of 3D Video*, May 2009, pp. 1–4.

[12] D. P. Bertsekas, *Nonlinear programming.* Belmont, MA, USA: Athena Scientific, 1999.

[13] Eur. Telecommun. Stand. Inst., "Universal mobile telecommunications system (UMTS): Multiplexing and channel coding (FDD)," *3GPP TS 125.212 Ver. 3.4.0*, pp. 14–20, Sep. 23, 2000.