JOINT SOURCE-CHANNEL CODING OF 3D VIDEO USING MULTIVIEW CODING

Arash Vosoughi, Vanessa Testoni, Pamela Cosman, and Laurence Milstein

ECE, University of California, San Diego

ABSTRACT

We consider the joint source-channel coding problem of a 3D video transmitted over an AWGN channel. The goal is to minimize the total number of bits, which is the sum of the number of source bits and the number of forward error correction bits, under two constraints: the quality of the primary view and the quality of the secondary view must be greater than or equal to a predetermined threshold at the receiver. The quality is measured in terms of the expected PSNR of an entire decoded group of pictures. A MVC (multiview coding) encoder is used as the source encoder, and rate compatible punctured turbo codes are utilized for protection of the encoded 3D video over the noisy channel. Equal error protection and unequal error protection are compared for various 3D video sequences and noise levels.

Index Terms— Joint source-channel coding, multiview video coding, turbo codes, unequal error protection

1. INTRODUCTION

When channel errors are present, the search for the optimal point in allocating bits between source and channel coding is one type of joint source channel coding (JSCC). The tradeoff between source coding accuracy and channel error protection of single-view video sequences in error-prone channels is a well-studied area. The work in [1] presents a comprehensive review on this topic while the work in [2] applies JSCC specifically for video transmission over additive white Gaussian noise (AWGN) channels using rate compatible punctured convolutional (RCPC) codes. The optimal point found by JSCC varies over different AWGN channel signal to noise ratios (SNRs).

The channel error protection provided by JSCC can be improved if unequal error protection (UEP) techniques are added to the problem formulation. UEP can be achieved by employing different channel code rates for each video packet. The distortion of the reconstructed video should be reduced when compared to the reconstructed video protected with equal error protection (EEP), where all video packets are coded with the same channel code rate. The channel code rate chosen to protect each packet depends on the estimation of the distortion produced by the packet loss. The distortion estimation can rely on traditional quality metrics, such as mean-squared error (MSE), or on metrics based on human visual perception, such as the packet loss visibility model presented in [3]. UEP is applied in [3] for optimizing the channel code rate for transmission of pre-encoded single-view video sequences over AWGN channels. The packet code allocation problem is designed as an integer programming problem and solved using variations of the branch-and-bound method.

In this paper, we propose a combined JSCC/UEP scheme for video transmission over AWGN channels. The main difference from the cited works is that we address 3D video coding with the multiview video coding (MVC) standard instead of single-view video coding. A second different is that rate compatible punctured turbo (RCPT) codes are employed instead of RCPC codes. The proposed JSCC scheme also differs from prior work in the way that it is formulated. For many image/video JSCC problems, a typical optimization approach is to fix a total rate of B bits and then determine the optimal division of B between source coding and FEC where the objective function could be the PSNR to be maximized. An example of this type of optimization can be found in [4], where MVC is used for source coding, Raptor codes are used as FEC, and a weighted average of the PSNR of the left view and the PSNR of the right view is used as the quality metric for stereo video. Formulating the optimization in this way is problematic because although the PSNR for the left view is well defined, as is the PSNR for the right view, there is not yet any well accepted way to quantify the distortion or PSNR of the combined stereo video. A quality metric for 3D stereo video is an unsolved and difficult problem [5]. Maximizing the average PSNR subject to a rate constraint would consider that left/right PSNRs of 20/40 and of 30/30 produce equivalent average PSNRs, although the subjective visual quality might be very different. Indeed, if the rate constraint were such that the one of the two views could have distortion driven to near zero, so the PSNR approaches infinity, the distortion of the other view could be arbitrarily bad and yet the average PSNR might be maximized. Our alternative approach to the optimization is to fix the distortion or PSNR of each view to some level, and then attempt to minimize the number of bits required to achieve it. Putting the distortion in the constraint, rather than in the objective function to be minimized, allows one to choose two separate constraints (one for each view).

The performance and transmission of MVC in error-prone channels has been studied from several aspects in [6, 7, 8]. Some of the works in the area of multiview streaming optimization, as in [6], propose end-to-end distortion models taking into account estimated packet loss probabilities for multiview video packets, but do not include channel error protection schemes. The work in [7] has the same characteristics, but includes a form of UEP by simply setting a smaller packet loss rate for the packets in the base view as well as the packets in the first 20 frames of the other views. Another work [8] that studied the transmission of multiview video sequences over error-prone channels considered UEP through a selective packet discard mechanism. In this paper, source coding using MVC, channel coding using RCPT, and unequal error protection are considered for 3D video sequences. The MVC *base view* is denoted here as the *primary view*, and corresponds to the left-eye stereo view. The right-eye stereo view is denoted here as the *secondary view*, and corresponds to the MVC *enhancement view*.

Due to the inter-view dependencies exploited by MVC, the expected end-to-end estimate of the secondary view distortion also takes into account the distortion generated by primary view packet losses. The packet channel code rate allocation scheme in [3] is extended to include the choice of the optimum packet quantization parameters (QPs) for both views. A particular characteristic of our scheme is that it is subjected to quality constraints for both views, instead of the usual bit-rate

This research was supported by the Intel/Cisco Video Aware Wireless Networks (VAWN) program, by InterDigital, Inc., and by the National Science Foundation under grant number CFF-1160832.

constraints. Therefore, our goal is to minimize the total bit-rate, composed of source bits and FEC, while both reconstructed views achieve a predetermined minimum PSNR value.

The details of the JSCC/UEP scheme and of the end-to-end distortion model are presented in Section 2. The simulation setup and the results obtained with the JSCC/UEP approach when compared with the JSCC/EEP approach are shown in Section 3. Finally, Section 4 presents the conclusions.

2. PROBLEM FORMULATION

2.1. Modeling the End-to-End Distortion

In this section, we model the end-to-end distortion of a GOP of the primary and secondary views of an MVC-encoded 3D video sent over a noisy channel. A model is used to obtain a UEP solution, which enables us to assign different code rates to different packets such that the end-to-end distortions predicted by the model satisfy the quality constraints. We also emphasize that: 1) in Section 3, we verify by simulating many channel realizations that the end-to-end distortion constraints are met when the UEP solution is used, 2) because the UEP from the model may not be the best UEP solution, our resulting gain given in Section 3 is a lower bound on the possible UEP gain over EEP, 3) throughout the paper, by EEP we mean that all of the packets are protected by the same code rate, and this code rate is determined by simulation and not by the model.

Let $f^{(v)}$ represent the original pixel values of a GOP of view v, and $\widehat{f^{(v)}}$ be the corresponding reconstruction values at the encoder, where $v = 1$ represents the primary view and $v = 2$ represents the secondary view. We denote the reconstructed pixel values of the decoded GOP at the decoder as $\widetilde{f^{(v)}}$. The distortion of the GOP is the sum of distortions of all of the pixels in the GOP, and assuming that the source quantization distortion and the channel distortion are uncorrelated ([4], [9], [10]), the expected distortion of the GOP of view v can be written as:

$$\begin{aligned} D^{(v)} &= E\{\text{cmse}(f^{(v)}, \widehat{f^{(v)}})\} + E\{\text{cmse}(\widehat{f^{(v)}}, \widetilde{f^{(v)}})\} \\ &= D^{(v)}_{Src} + D^{(v)}_{Loss}, \end{aligned} \quad (1)$$

where $\text{cmse}(x, y)$ is the cumulative mean squared error (CMSE) between the pixels of GOP x and GOP y, $D^{(v)}_{Src}$ represents the source distortion over the entire GOP of view v, and $D^{(v)}_{Loss}$ denotes the distortion introduced by the channel due to the packet losses. In this paper, the precise value of $D^{(v)}_{Src}$ is computed at the encoder and used in the simulations. To compute $D^{(v)}_{Loss}$, we assume that the MVC-encoded 3D video is packetized such that each row of macroblocks of the encoded video is encapsulated as a separate packet. This packetization is included in the H.264/AVC standard for error resiliency [11]. Now, we assume that the errors propagated within the primary view due to lost packets in the primary view do not affect each other. We also assume that the errors propagated in the secondary view, which may be due to lost packets either in the primary view or in the secondary view, do not affect each other. The two assumptions above are reasonable since there are only a few packets which are lost in transmission.

With these assumptions, the CMSE contribution of the individual packets to the CMSE of the entire GOP of either the primary view or secondary view would be additive. To compute $D^{(v)}_{Loss}$, we assume that the mth packet of view v is lost with probability $1 - \left(1 - p_b(r^{(v)}_m, SNR)\right)^{S^{(v)}_m}$, where $S^{(v)}_m$ indicates the size of the packet in bits, $r^{(v)}_m$ is the RCPT code rate allocated to the packet, and p_b represents the bit error probability after the channel decoder which can be computed by simulation. In the following, $D^{(v')}_{m,v}$ denotes the CMSE contribution of the mth packet of view v to the CMSE of the entire GOP of view v', and, $\widetilde{f^{(v')}_{m,v}}$ represents the reconstructed GOP of view v' at the decoder when the mth packet of view v is lost.

Let us first consider the primary view. The CMSE contribution of the ith packet of the primary view to the CMSE of the entire GOP of the primary view is zero if the packet is not lost, and is equal to $D^{(1)}_{i,1}(q_1)$ if the packet is lost, where q_1 is the quantization parameter used to encode the GOP. Thus, following (1), we have:

$$\begin{aligned} D^{(1)}(q_1, r^{(1)}_1, \ldots, r^{(1)}_K, SNR) &= D^{(1)}_{Src}(q_1) \\ &+ \sum_{i=1}^{K} \left(1 - \left(1 - p_b(r^{(1)}_i, SNR)\right)^{S^{(1)}_i(q_1)}\right) D^{(1)}_{i,1}(q_1), \end{aligned} \quad (2)$$

where K is the number of primary view packets in the GOP (which is the same as the number of secondary view packets in the GOP), and $D^{(1)}_{i,1}(q_1)$ is equal to $\text{cmse}(\widehat{f^{(1)}}(q_1), \widetilde{f^{(1)}_{i,1}}(q_1))$.

The distortion generated in the secondary view can be formulated in a similar manner. However, since the error due to a lost packet in the primary view propagates in both of the primary and secondary views, for the secondary view, the CMSE contribution of lost primary packets should be considered as well as the CMSE contribution of lost secondary packets. Therefore, with the assumption that error propagating from one lost packet does not interact with that from another lost packet, we have:

$$\begin{aligned} D^{(2)}(q_1, q_2, r^{(1)}_1, \ldots, r^{(1)}_K, r^{(2)}_1, \ldots, r^{(2)}_K, SNR) &= D^{(2)}_{Src}(q_1, q_2) \\ &+ \sum_{i=1}^{K} \left(1 - \left(1 - p_b(r^{(1)}_i, SNR)\right)^{S^{(1)}_i(q_1)}\right) D^{(2)}_{i,1}(q_1, q_2) \\ &+ \sum_{j=1}^{K} \left(1 - \left(1 - p_b(r^{(2)}_j, SNR)\right)^{S^{(2)}_j(q_1, q_2)}\right) D^{(2)}_{j,2}(q_1, q_2), \end{aligned} \quad (3)$$

where $D^{(2)}_{i,1}(q_1, q_2)$ is equal to $\text{cmse}(\widehat{f^{(2)}}(q_1, q_2), \widetilde{f^{(2)}_{i,1}}(q_1, q_2))$, and $D^{(2)}_{j,2}(q_1, q_2)$ is equal to $\text{cmse}(\widehat{f^{(2)}}(q_1, q_2), \widetilde{f^{(2)}_{j,2}}(q_1, q_2))$. The quantities $D^{(1)}_{i,1}(q_1)$, $D^{(2)}_{i,1}(q_1, q_2)$ and $D^{(2)}_{j,2}(q_1, q_2)$ are computed precisely at the encoder, and used in the simulations.

2.2. JSCC of 3D Video Using Integer Optimization

The objective of our JSCC problem is to minimize the total number of bits, which is the sum of the number of source bits and the number of FEC bits of both the primary and secondary views. Therefore, the objective function is formulated as:

$$\min_{\substack{q_1 \in QP_1 \\ q_2 \in QP_2 \\ r^{(1)}_1, \ldots, r^{(1)}_K \in R \\ r^{(2)}_1, \ldots, r^{(2)}_K \in R}} \left(\sum_{i=1}^{K} \frac{S^{(1)}_i(q_1)}{r^{(1)}_i} + \sum_{j=1}^{K} \frac{S^{(2)}_j(q_1, q_2)}{r^{(2)}_j} \right), \quad (4)$$

where $R = \{R_1, R_2, \ldots, R_N\}$ is the set of available RCPT code rates, and QP_1 and QP_2 are the quantization parameter sets. The optimization is done over a primary view GOP and its corresponding secondary view GOP at the same time. Quantization parameters q_1 and q_2 are applied for all blocks of the primary view and secondary view GOPs. In minimizing the objective function (4), two constraints must be satisfied: the expected distortion of the primary view and the expected distortion

of the secondary view must be smaller than or equal to a predetermined threshold T at the receiver. Using (2) and (3), this can be expressed as:

$$D^{(1)}(q_1, r_1^{(1)}, \ldots, r_K^{(1)}, SNR) \leq T$$
$$D^{(2)}(q_1, q_2, r_1^{(1)}, \ldots, r_K^{(1)}, r_1^{(2)}, \ldots, r_K^{(2)}, SNR) \leq T. \quad (5)$$

In the optimization problem introduced in (4) and (5), different code rates are typically assigned to different packets. The assignment depends on the size of the packets at the output of the source encoder and the distortion they generate if they are lost in transmission. To solve this optimization problem, we search over a grid of quantization parameters q_1 and q_2. For each (q_1, q_2) we find the optimum code rates which minimize the total number of bits and meet constraint (5). The final solution is obtained as a quantization pair (q_1, q_2) and a set of code rates, which together produce the smallest total number of bits. This problem is a nonlinear integer programming problem, which can be solved by the branch-and-bound (BnB) method. To employ BnB (which is based on binary variables) we transform the original integer optimization variables into binary variables as suggested in [12]. Each variable $r_i^{(1)}$ is transformed to N binary variables $x_{i,s}^{(1)}$ $(1 \leq s \leq N)$ and each variable $r_j^{(2)}$ is also transformed to N binary variables $y_{j,t}^{(2)}$ $(1 \leq t \leq N)$, where x and y take values from the set $\{0, 1\}$. $r_i^{(1)}$ is then substituted with $\sum_{s=1}^{N} x_{i,s}^{(1)} R_s$ and $r_j^{(2)}$ is substituted with $\sum_{t=1}^{N} y_{j,t}^{(1)} R_t$ in (4) and (5). With these transformations, $2K$ equality constraints are considered along with the inequalities given in (5), which are:

$$\sum_{s=1}^{N} x_{i,s}^{(1)} = 1, \quad 1 \leq i \leq K$$
$$\sum_{t=1}^{N} y_{j,t}^{(1)} = 1, \quad 1 \leq j \leq K. \quad (6)$$

Fig. 1 shows the total number of bits obtained by the EEP approach and the proposed UEP approach for one GOP of 'Rena', where SNR = 3 dB. The UEP solution is obtained by solving the optimization problem given in (4) and (5). The EEP solution is obtained by exhaustive simulation of all possible EEP rates. It is observed that the minimum number of bits obtained by the EEP and UEP correspond to the quantization parameter pairs $(30, 29)$ and $(29, 29)$, respectively. The UEP leads to 12.9% bit savings compared to EEP. Note that the curves cannot be continued to the right because with higher quantization parameters, the quality constraint is not met.

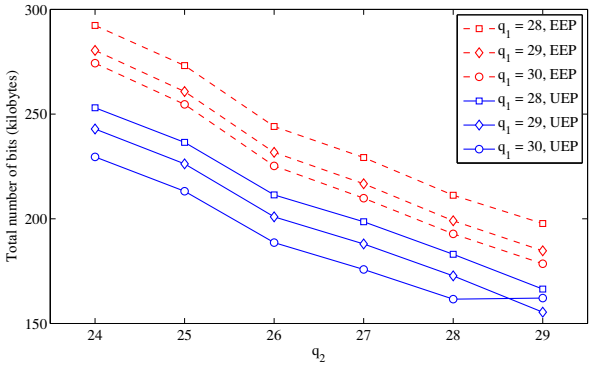

Fig. 1: Total number of bits of 10 frames of 'Rena' versus q_1 and q_2 for different protection approaches. Channel is AWGN and SNR = 3 dB.

3. SIMULATION RESULTS

Simulation results for the AWGN channel are given in this section. BPSK modulation is employed. Video sequences 'Exit', 'Rena', and 'Flamenco' with resolution 640×480, and 'Ballet' with resolution 1024×768 are used. We used the JM 18.2 reference software for encoding the tested stereo video sequences, where each row of macro blocks of either the primary view or secondary view is encoded as a slice. We used the decoder of the JMVC 8.2 reference software for decoding the MVC bit stream, where we implemented linear interpolation for error concealment of lost I slices, and slice copy for lost P slices such that a lost P slice is concealed from its reference frame in the same view. The GOP structures of the primary view and the secondary view are IPPP... and PPPP..., respectively, and the GOP size is 10 frames, which is in accordance with small cyclic-Intra coded period as required in [13]. We used UMTS turbo codes for channel coding. The UMTS turbo encoder is composed of two recursive systematic convolutional encoders with constraint length 4, which are concatenated in parallel [14]. The feedforward and feedback generators are 15 and 13 respectively, both in octal. The mother code rate of the RCPT code is $\frac{1}{3}$, the puncturing period used is $P = 8$, and the set of available rates of the RCPT code is $\{\frac{4}{5}, \frac{8}{11}, \frac{2}{3}, \frac{4}{7}, \frac{8}{15}, \frac{1}{2}, \frac{8}{17}, \frac{4}{9}, \frac{8}{19}\}$. An iterative, soft-input/soft-output (SISO) decoding algorithm is used for turbo decoding. We considered eight iterations to compute the decoded BERs. The threshold T in (5) is set such that the PSNR of both views is at least 40 dB at the receiver. In the simulations, the quantization parameter set QP_1 is a subset of $\{22, 23, \ldots, 29, 30\}$ and the quantization parameter set QP_2 is a subset of $\{19, 20, \ldots, 28, 30\}$. Note that, depending on a specific video sequence, there may be some quantization pairs (q_1, q_2) for which the PSNR of the noise free encoded 3D video does not satisfy the 40 dB constraints and so they are excluded from consideration.

We compare the total number of bits obtained by the proposed UEP approach, which is determined by solving the optimization problem given in (4) and (5), to that of the EEP approach, which is obtained by simulation. The relative difference in the total number of bits is defined as:

$$e = \frac{\#bits^{(\text{EEP})} - \#bits^{(\text{UEP})}}{\#bits^{(\text{EEP})}} \times 100\%. \quad (7)$$

Table 1 shows the results for 100 frames of each video sequence and different noise levels. The UEP approach reduces the total number of bits up to 13.4%, 12.5%, 12.6%, and 13.5% for 'Rena', 'Exit', 'Flamenco', and 'Ballet', respectively. The average gain of UEP compared

Table 1: Relative difference between the total number of bits obtained by the EEP and UEP for AWGN channel and 100 frames of each tested video sequence.

video sequence	SNR	e
Rena (640×480)	3 dB	13.4%
	4 dB	11.3%
	5 dB	10.8%
Exit (640×480)	3 dB	12.5%
	4 dB	11.2%
	5 dB	10.6%
Flamenco (640×480)	3 dB	12.6%
	4 dB	10.1%
	5 dB	8.6%
Ballet (1024×768)	3 dB	13.5%
	4 dB	10.9%
	5 dB	10.3%

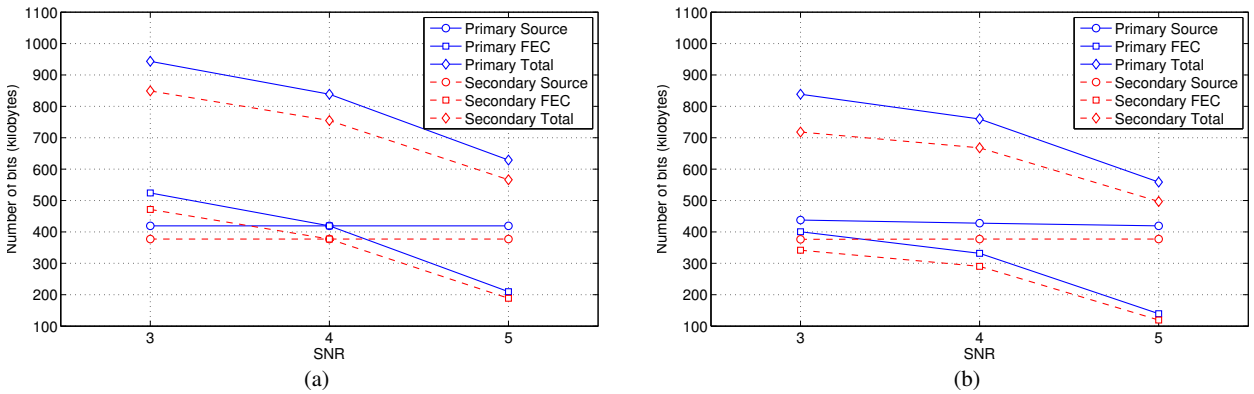

Fig. 2: (a) Number of source bits, FEC bits, and total bits obtained by EEP via simulation, (b) number of source bits, FEC bits, and total bits obtained by UEP. 100 frames of 'Rena' are used.

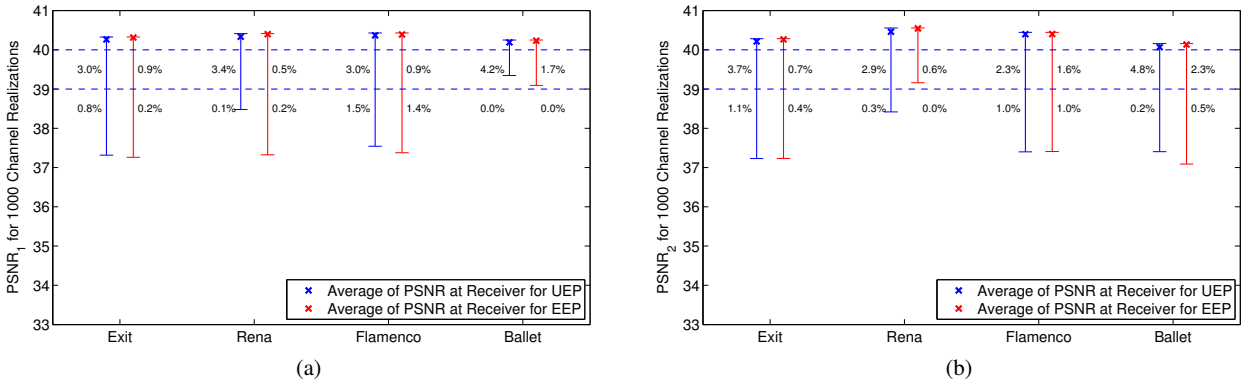

Fig. 3: (a), (b) PSNR range of the primary view ($PSNR_1$) and PSNR range of the secondary view ($PSNR_2$) at the receiver for 1000 AWGN channel realizations and the tested 3D video sequences.

to EEP is 11.8, 11.4, 10.4, and 11.6 for 'Rena', 'Exit', 'Flamenco', and 'Ballet', respectively.

In Figs. 2 (a) and (b), the EEP and UEP approaches are compared in terms of the number of source bits, number of FEC bits, and the number of total bits they require for different SNR values such that the quality constraints at the receiver are satisfied. As expected, the number of required FEC bits of both views decreases for higher SNR values. The UEP and EEP performance becomes the same for SNR values higher than the ones shown here, because these turbo codes are sufficiently powerful that, for higher SNRs, the quality constraints can be met using the weakest code equally for all the data.

The UEP solution obtained by the model was validated via simulation to see if the quality constraints are met for realistic channel realizations. Fig. 3(a) and Fig. 3(b) show the range of PSNRs of the two views at the receiver for the tested 3D video sequences and 1000 AWGN channel realizations with SNR = 4 dB. The numbers given next to each bar indicate the percentage of obtained PSNR values between 39 dB and 40 dB and less than or equal to 39 dB. From these figures, we see that the average PSNRs always meet the 40 dB constraint. Only a small percentage of the obtained PSNRs are between 40 dB and 39 dB, and the number of obtained PSNR values below 39 dB is negligible.

4. CONCLUSIONS

We addressed the JSCC problem of a 3D video sent over an AWGN channel with the goal of minimizing the total number of transmitted bits while subject to video quality constraints. The unequal error protection optimization approach proposed here proved to be efficient at achieving this goal when compared to equal error protection for an AWGN channel. Average gains vary from 10.4% to 11.8% (according to the SNR) when the UEP approach is compared to the EEP approach.

5. REFERENCES

[1] R. E. Van Dyck and D. J. Miller, "Transport of wireless video using separate, concatenated, and joint source-channel coding," *Proceedings of the IEEE*, pp. 1734–1750, 1999.

[2] M. Bystrom and J. W. Modestino, "Combined source-channel coding schemes for video transmission over an additive white gaussian noise channel," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 880–890, 2000.

[3] T.-L. Lin and P.C. Cosman, "Efficient optimal rcpc code rate allocation with packet discarding for pre-encoded compressed video," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 505 –508, may 2010.

[4] A. S. Tan, A. Aksay, G. B. Akar, and E. Arikan, "Rate-distortion optimization for stereoscopic video streaming with unequal error protection," *EURASIP J. Appl. Signal Process.*, vol. 2009, pp. 7:1–7:14, Jan. 2008.

[5] H.-T. Quan, P. Le Callet, and M. Barkowsky, "Video quality assessment: from 2D to 3D- challenges and future trends," in *Int. Conf. on Image Proc. (ICIP)*. IEEE, Sept. 2010, pp. 4025–4028.

[6] Y. Liu, Q. Huang, S. Ma, D. Zhao, and W. Gao, "RD-optimized interactive streaming of multi view video with multiple encodings," *Journal of Visual Communication and Image Representation*, pp. 523–532, 2010.

[7] Y. Zhou, C. Hou, W. Xiang, and F. Wu, "Channel distortion modeling for multi-view video transmission over packet-switched networks," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1679 – 1692, 2011.

[8] D. M. Bento and J. M. Monteiro, "A QoS solution for three-dimensional full-HD H.264/MVC video transmission over IP networks," *IEEE Iberian Conference on Information Systems and Technologies*, pp. 1–6, 2012.

[9] Y. Zhou, C. Hou, W. Xiang, and F. Wu, "Channel distortion modeling for multi-view video transmission over packet-switched networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 11, pp. 1679 –1692, nov. 2011.

[10] Y. Wang, Z. Wu, and J. M. Boyce, "Modeling of transmission-loss-induced distortion in decoded video," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 16, no. 6, Sept. 2006.

[11] A. Vetro, T. Wiegand, and G.J. Sullivan, "Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard," IEEE, Apr. 2011, pp. 626–642.

[12] D. P. Bertsekas, *Nonlinear Programming, 2nd Ed.*, Athena Scientific.

[13] ISO/IEC JTC1/SC29/WG11 & ITU-T SG16 Q.6, "Common conditions for MVC," *JVT-U211*, 2006.

[14] European Telecommunications Standards Institute, "Universal mobile telecommunications system (UMTS): Multiplexing and channel coding (FDD)," *3GPP TS 125.212 version 3.4.0*, pp. 14–20, Sept. 23 2000.