OPTIMAL BANDWIDTH ALLOCATION FOR SCALABLE H.264 VIDEO TRANSMISSION OVER MIMO SYSTEMS

Mohammad K. Jubran, Manu Bansal, Rohan Grover, Lisimachos P. Kondi
Department of Electrical Engineering,
State University of New York at Buffalo,
Buffalo, NY 14260.
Email: {mkjubran, mbansal, rgrover, lkondi}@eng.buffalo.edu

ABSTRACT

In this paper, we propose an optimal strategy for the transmission of scalable video over packet-based multiple-input multiple-output (MIMO) systems. The latest scalable H.264 codec is used, which provides combined temporal, quality and spatial scalability. In this work, we propose different error concealment schemes to handle packet losses at the decoder. At the encoder, we have developed a method for the estimation of the video distortion at the receiver for given channel conditions. We show the performance of our distortion estimation algorithm in comparison with simulated video transmission over wireless channels with packet errors. In the proposed system, we use a MIMO system with orthogonal space-time block codes (O-STBC) that provides spatial diversity and guarantees independent transmission of different symbols within the block code. Rate-compatible turbo codes (RCPT) are used for unequal error protection of the scalable layers. In the proposed constrained bandwidth allocation framework, we use the estimated decoder distortion to optimally select the application layer parameters, i.e. quantization parameter (QP) and group of pictures (GOP) size, and physical layer parameters, i.e. RCPT and symbol constellation. Also, the simulation results show the substantial performance gain by using different symbol constellations across the scalable layers as compared to a fixed constellation.

I. INTRODUCTION

The H.264/AVC standard has already been shown to provide superior compression efficiency and error-resilient transmission over varied networks [1], [2] and the very recently proposed scalable extension of H.264/AVC inherits its error-resilient network adaptation layer (NAL) structure and provides a combined scalability in the form of temporal scalability using a hierarchical prediction structure, fine granular quality scalability (FGS) using progressive refinement slices and spatial scalability using inter-layer prediction mechanisms [3], [4], [5]. This scalability can be exploited to improve the video transmission over error-prone wireless networks by protecting the different layers with unequal error protection (UEP). In this work, we explore temporal scalability and FGS, however the work can be extended to also include the spatial scalability as well.

We consider the efficient transmission of these scalable layers over packet-based wireless networks, with optimization of source coding, channel coding and physical layer parameters. For that to be possible, a good knowledge of the total end-to-end decoder distortion should be available at the encoder. Various decoder distortion estimation algorithms have been proposed in the literature. In [6] and [7], a recursive per-pixel based decoder distortion estimation algorithm, ROPE was proposed for non-scalable and scalable H.263+ codec, respectively, which is used for optimal mode selection for a given target rate. Also, in [8], Shen *et. al.* further modified the ROPE algorithm for different re-synchronization schemes for the transmission of non-scalable H.263 coded video over tandem channels. In our paper, we develop a method for the accurate estimation of the distortion of scalable H.264/AVC coded video at the receiver for given channel conditions and also propose different error concealment schemes to handle the packet losses. Our scalable decoder distortion estimation (SDDE) algorithm takes into account loss of both temporal and SNR scalable layers as well as error concealment at the decoder.

Diversity techniques, including spatial, time and frequency domain diversity, have been proven to help overcome the degradations due to wireless channels (noise, fading etc.) by providing the receiver with multiple replicas of the transmitted signal over different channels. As one of the diversity techniques, space-time coding (STC) over multiple antenna systems has been studied extensively [9], [10]. Space-time block codes (STBC),

which were first proposed by Alamouti [9] and later generalized by Tarokh *et. al.* [10], are one of the STC techniques for broadband wireless communications. The orthogonal STBC (O-STBC) used here for video transmission over the MIMO system guarantee independent transmission and low-complexity maximum-likelihood (ML) decoding of each symbol in a given block code. This enables us to independently choose the elements of the codeword from different constellations.

In only a few publications such as [11], [12], [13], wireless video transmission using STC has been studied. In [11], a joint source-channel matching framework for image transmission using the SPIHT encoder over an OFDM system with space-time block codes is proposed. Zhao *et. al.* in [12], proposed progressive video transmission over a space-time differentially coded OFDM system. They proposed an UEP structure with optimal rate and power allocation among multiple layers. However, in all the above-mentioned work, the orthogonal structure of STBC codes has not been exploited by independent transmission of the layered video over different symbols of the STBC code modulated with different constellations. In [14], an approach for using the scalable H.264 with unequal erasure protection (UXP) over wireless IP networks has been proposed.

In this paper, we propose a system that integrates video coding with combined scalability, forward error correction (FEC) through unequal channel coding, modulation schemes and spatial diversity for wireless video transmission. Temporal and quality scalable layers are obtained using a scalable H.264 codec and are unequally protected using rate-compatible punctured turbo (RCPT) [15] codes with cyclic redundancy check (CRC) [16] error detection. The channel coded layers are then modulated and encoded using O-STBC for transmission over multiple antennas. We address the problem of minimization of the expected end-to-end distortion by optimally selecting source coding parameters: the quantization parameter (QP) and the group of pictures (GOP) size, and the physical layer parameters: RCPT channel coding rate and the symbol constellation choice for the MIMO transmission. The optimization is constrained on the total available bandwidth for transmission. The accurate estimation of the decoder distortion using the SDDE algorithm plays a key role in this optimization problem.

II. SCALABLE H.264 CODEC AND DECODER DISTORTION ESTIMATION

The latest scalable extension of H.264/AVC is based on a hierarchical prediction structure as shown in Figure

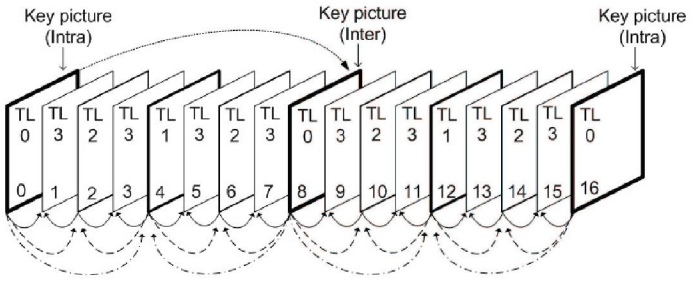


Fig. 1. Hierarchical prediction structure for scalable H.264 for GOP size = 8, two GOPs are shown.

1. The first picture of the video sequence is always intra-coded and is a key picture. A GOP consists of a key picture and all other pictures temporally located between the key picture and the previously encoded key picture. The key pictures are either encoded in intra or inter mode, with only previously encoded key pictures as the reference pictures. These key pictures collectively form the lowest temporal resolution of the video sequence and are called temporal level zero (TL0). The other pictures encoded in each GOP define different temporal levels (TL1, TL2, so on) and always use the pictures from the lower temporal levels as the reference pictures. Each of these pictures is represented by a non-scalable base layer (FGS0) that includes the corresponding motion and an approximation of the intra and residual data, and zero or more quality scalable enhancement (FGS) layers. From here on in this paper, the base and enhancement layers will be referred to as the SNR scalable layers where the base layer of each frame is associated with a particular temporal level. In our work, we limit ourselves to temporal and SNR scalability such that priority for the base layer (FGS0) of each temporal level increases from the lowest to the highest temporal level. And then each FGS layer for all the frames is considered as a single layer.

In our proposed system, for efficient transmission of video over wireless networks, we have to optimally select various system parameters at the transmitter. To do so, at the encoder we should have a good knowledge of the total decoded video distortion (due to source coding and channel errors). For this purpose, we have developed a method for the accurate estimation of the video distortion at the receiver for given channel conditions.

In the SDDE algorithm, we calculate the distortion on a per-pixel basis, and the distortion of each frame is the summation of the distortion calculated for all the pixels in the corresponding frame. Each layer of each frame is packetized into constant size packets, which are transmitted over lossy wireless networks. At the receiver,

any unrecoverable errors in each packet would result in dropping (erasing) of that packet and hence would mean loss of the layer (of that particular frame) to which the packet belongs. In this system, we guarantee that the base layers of all the key pictures are received error free. It should be emphasized here that the scalable H.264 encoding and decoding are done on a GOP basis (using the hierarchical structure) which makes it possible to use the frames within a GOP for error concealment purposes. In the event of losing a frame, temporal error concealment at the decoder is applied. We consider the following four error concealment schemes:

- Scheme 1: The lost frame is replaced by the previous frame in the decreasing sequential order, e.g. in a GOP if frame f_n is lost, it is concealed using frame f_{n-1} given that it was received error free, otherwise f_{n-2} is used for concealment and so on, till the start of GOP is reached.
- Scheme 2: The lost frame is replaced by the nearest available frame in the decreasing as well as increasing sequential order. We start towards the GOP end closer to the frame being concealed, e.g. in a GOP of eight frames, if frame f_6 is lost, the order in which the frames of this GOP are used for concealment is f_7, f_5 and finally f_8.
- Scheme 3: The lost frame is replaced by the previous reconstructed frame in the sequential order that leads faster towards the GOP end, and only using the frames from lower or same temporal levels, e.g. in a GOP of eight frames, if frame f_5 is lost, the order in which the frames are used for concealment is f_6 and then f_8. However, for frame f_3 the concealment order is f_2 and f_0. For the frame in the center of the GOP (like f_4), the key picture at the start of the GOP is used for concealment.
- Scheme 4: The lost frame is replaced by the nearest available frame in the decreasing as well as increasing sequential order from only lower or same temporal levels. We start towards the frames that have a temporal level closer to the temporal level of the lost frame, e.g. in a GOP of eight frames, if frame f_6 is lost, the order in which the frames are used for concealment is f_4 and then f_8. As in Scheme 3, the center picture of the GOP is concealed using the starting key picture.

Next we will show the derivation for SDDE, in which we will consider a base layer and two FGS layers. The same algorithm can simply be generalized to any number of FGS layers. We assume the frames are lexicographically ordered. Let f_n^i denote the original value of pixel i in frame n and $\hat{f}_n^i$ denote its encoder reconstruction. The reconstructed pixel value at the decoder is denoted by $\tilde{f}_n^i$. The mean square error for this pixel is

$$d_n^i = \mathrm{E}\left\{\left(f_n^i - \tilde{f}_n^i\right)^2\right\} = \left(f_n^i\right)^2 - 2f_n^i \mathrm{E}\left\{\tilde{f}_n^i\right\} + \mathrm{E}\left\{\left(\tilde{f}_n^i\right)^2\right\} \quad (1)$$

where d_n^i is the distortion per pixel. As mentioned earlier, the base layer of all the key pictures are guaranteed to be received error free, the s^{th} moment of the i^{th} pixel of the key pictures n is calculated as follows:

$$\begin{aligned}\mathrm{E}\left\{\left(\tilde{f}_n^i\right)^s\right\} &= P_{nE1}\left(\hat{f}_{nB}^i\right)^s + (1-P_{nE1})\,P_{nE2}\left(\hat{f}_{n(B+E1)}^i\right)^s \\ &\quad + (1-P_{nE1})\,(1-P_{nE2})\left(\hat{f}_{n(B+E1+E2)}^i\right)^s\end{aligned} \quad (2)$$

where $\hat{f}_{nB}^i$, $\hat{f}_{n(B+E1)}^i$, $\hat{f}_{n(B+E1+E2)}^i$ are the reconstructed pixel values at the encoder of only the base layer, the base along with the first FGS layer and the base layer with both of the FGS layers of frame n, respectively. P_{nE1} and P_{nE2} are the probabilities of losing the first and the second FGS layer of frame n, respectively.

For all the frames except the key pictures of a GOP, let us denote $\hat{f}_{nB\_uv}^i$ as the i^{th} pixel value of the base layer of frame n reconstructed at the encoder. Frames $u(<n)$ and $v(>n)$ are the reference pictures used in the hierarchical prediction structure for the reconstruction of frame n. In the decoding process of scalable H.264, the frames of each GOP are decoded in the order starting from the lowest to the highest temporal level. At the decoder,

- If frame u is not available as the reference picture for frame n (where frame n does not belong to the highest temporal level), then frame u' is selected as the new reference picture such that $u' < n$ and $TL(u') \leq TL(n)$ where $TL(.)$ is the temporal level to which the corresponding frame belongs. For the frames in the highest temporal level, $u' < n$ and $TL(u')$ is strictly less than $TL(n)$. Let us define $\mathbf{L}_n$ as the set consisting of frame u and all the possible choices of u' for frame n.
- If frame v is not available as the reference picture for frame n, then frame v' is selected as the new reference picture such that $v' > n$ and $TL(v') < TL(n)$. In this case, we define $\mathbf{R}_n$ as the set consisting of frame v and all the possible choices of v' for frame n.

The s^{th} moment of the i^{th} pixel of frame n when at least the base layer is received correctly is defined as:

$$\mathrm{E}\left\{\left(\tilde{f}_n^i\left(\mathbf{L}_n,\mathbf{R}_n\right)\right)^s\right\}=\sum_{j=1}^{|\mathbf{L}_n|}\sum_{k=1}^{|\mathbf{R}_n|}\left(1-P_{\mathbf{L}_n(j)}\right)\left(1-P_{\mathbf{R}_n(k)}\right)\prod_{c=1}^{j-1}P_{\mathbf{L}_n(c)}\prod_{d=1}^{k-1}P_{\mathbf{R}_n(d)}\mathrm{E}\left\{\left(\tilde{f}^i_{n\_\mathbf{L}_n(j)\mathbf{R}_n(k)}\right)^s\right\} \tag{3}$$

where,

$$\begin{aligned}\mathrm{E}\left\{\left(\tilde{f}^i_{n\_\mathbf{L}_n(j)\mathbf{R}_n(k)}\right)^s\right\}&=P_{nE1}\left(\hat{f}^i_{nB\_\mathbf{L}_n(j)\mathbf{R}_n(k)}\right)^s\\&+P_{nE2}\left(1-P_{nE1}\right)\left(\hat{f}^i_{n(B+E1)\_\mathbf{L}_n(j)\mathbf{R}_n(k)}\right)^s\\&+\left(1-P_{nE2}\right)\left(1-P_{nE1}\right)\left(\hat{f}^i_{n(B+E1+E2)\_\mathbf{L}_n(j)\mathbf{R}_n(k)}\right)^s\end{aligned} \tag{4}$$

where, $P_{\mathbf{L}_n(j)}$ and $P_{\mathbf{R}_n(k)}$ are the probabilities of losing the base layer of the reference frames j and k from the sets $\mathbf{L}_n$ and $\mathbf{R}_n$, respectively.

Now to get the distortion per-pixel after error concealment, we will define a set $\mathbf{Q} = \{f_n, f_{q1}, f_{q2}, f_{q3}, ..., f_{GOPend}\}$, where f_n is the frame to be concealed, f_{q1} is the first frame, f_{q2} is the second frame to be used for concealment of f_n, and so on till one of the GOP ends is reached. The s^{th} moment of the i^{th} pixel using the set $\mathbf{Q}$ is defined as $\mathrm{E}\left\{\left(\tilde{f}_n^i\right)^s\right\}$,

$$\begin{aligned}&\mathrm{E}\left\{\left(\tilde{f}_n^i\right)^s\right\}=\left(1-P_n\right)\mathrm{E}\left\{\left(\tilde{f}_n^i\left(\mathbf{L}_n,\mathbf{R}_n\right)\right)^s\right\}\\&+P_n\left(1-P_{q1}\right)\mathrm{E}\left\{\left(\tilde{f}^i_{q1}\left(\mathbf{L}_{q1}\setminus\{f_n\},\mathbf{R}_{q1}\setminus\{f_n\}\right)\right)^s\right\}\\&+P_nP_{q1}\left(1-P_{q2}\right)\mathrm{E}\left\{\left(\tilde{f}^i_{q2}\left(\mathbf{L}_{q2}\setminus\{f_n,f_{q1}\},\mathbf{R}_{q2}\setminus\{f_n,f_{q1}\}\right)\right)^s\right\}\\&+\ldots+P_n\prod_{z=1}^{|\mathbf{Q}|-2}P_{qz}\mathrm{E}\left\{\left(\tilde{f}^i_{GOPend}\right)^s\right\}\end{aligned} \tag{5}$$

where P_n and P_{qz} are the probabilities of losing the base layer of frame n and qz, respectively; $\mathbf{L}_n \setminus \{f_w\}$ is the set of all the reference frames $\mathbf{L}_n$ excluding frame f_w, and $\mathbf{R}_n \setminus \{f_w\}$ is the set of all the reference frames $\mathbf{R}_n$ excluding frame f_w. The mean square error in (1) is obtained by calculating the the 1^{st} and 2^{nd} moments of pixel i of frame n using (2), (3), (4) and (5).

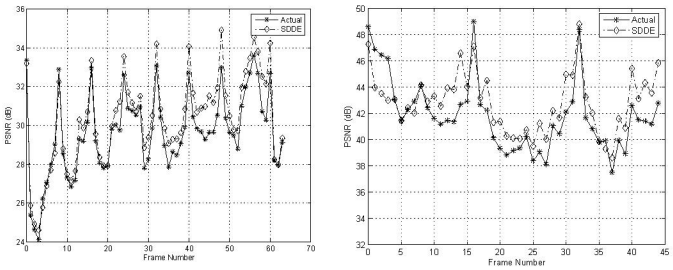
A. Performance analysis

The algorithm explained above is implemented by using the scalable H.264/AVC codec and its performance is evaluated by comparing it with the actual decoder distortion averaged over 200 channel realizations. We considered different video sequences (QCIF format) encoded at 30 fps. Figure 2(a) show the results for the "Foreman" video sequence encoded at QP = 40 and GOP size of eight frames. Each of these layers is considered to be affected with different loss rates as follows: $P_{TL0} = 0\%$, $P_{TL1} = 10\%$, $P_{TL2} = 20\%$, $P_{TL3} = 30\%$, $P_{E1} = 50\%$ and $P_{E2} = 60\%$ where P_{TLx} is the probability of losing the base layer of a frame that belongs to TLx, P_{E1} and P_{E2} are the probabilities of losing FGS1 and FGS2 of each frame, respectively. It is evident that the proposed SDDE algorithm provides an accurate estimate of the scalable H.264 decoder distortion at the encoder. Similar results are also shown in Figure 2(b) using the "Akiyo" video sequence encoded at QP = 25 and GOP =16. The packet error rates considered in this case are $P_{TL0} = 0\%$, $P_{TL1} = 10\%$, $P_{TL2} = 15\%$, $P_{TL3} = 20\%$, $P_{TL4} = 25\%$, $P_{E1} = 30\%$ and $P_{E2} = 40\%$. Also, comparison of average PSNRs (160 frames) for both the actual and SDDE are listed in the Table I. It can be observed that error concealment scheme 2 provides better PSNR values than those obtained by using other schemes and hence for the rest of the paper we will use scheme 2 for error concealment at the decoder.

TABLE I

THE ACTUAL AND SDDE AVERAGE PSNR VALUES (dB).

Error concealment	Foreman Actual	Foreman SDDE	Akiyo Actual	Akiyo SDDE
Scheme 1	30.27	31.12	42.22	43.68
Scheme 2	30.73	31.39	42.95	43.80
Scheme 3	29.64	30.34	41.65	42.63
Scheme 4	29.78	30.56	41.76	43.14



(a) (b)

Fig. 2. Comparison between the actual and estimated decoder PSNR, (a) "Foreman" sequence using error concealment scheme 3 and (b) "Akiyo" sequence using error concealment scheme 2.

III. SYSTEM DESCRIPTION

In our packet-based video transmission system, after the scalable encoding of the video, the base and FGS layers of each frame are divided into packets of constant size ($= \gamma$). These constant size source packets are then channel encoded using 16-bit CRC for error detection and rate-compatible punctured turbo (RCPT) codes for unequal error protection. These channel encoded packets are further encoded using O-STBC for transmission over MIMO wireless systems. A Rayleigh flat-fading channel with AWGN is considered between each transmitter and each receiver. At the receiver, maximum-likelihood (ML) decoding is used to detect the transmitted symbols which are then demodulated and channel decoded for error correction and detection. If a packet is not detected to be

error-free, the layer of the corresponding frame to which the packet belongs is dropped. All the error-free packets for each frame are buffered and then fed to the source decoder with error concealment for video reconstruction.

For the MIMO system, we consider M_t transmit and M_r receive antennas. We have considered two different MIMO systems with the same diversity gain ($M_t \times M_r$=4):

- MIMO System 1
  In MIMO System 1, we have taken $M_t = M_r = 2$ with the O-STBC design proposed by Alamouti [9] $\mathbf{G}_2(x_1, x_2)$ of rate 1, where x_1 and x_2 are the symbols that can be chosen from either same or different constellations, transmitted in $T = 2$ time slots.
  $$\mathbf{G}_2(x_1, x_2) = \begin{bmatrix} x_1 & x_2 \\ -x_2^* & x_1^* \end{bmatrix} \quad (6)$$
- MIMO System 2
  In this sytem, we have considered $M_t = 4$ and $M_r = 1$, and the O-STBC design proposed by Tarokh *et. al* [10], $\mathbf{G}_4(x_1, x_2, x_3)$ of rate $3/4$ is used, where x_1, x_2 and x_3 are the symbols that can be chosen from either same or different constellations, transmitted in $T = 4$ time slots.
  $$\mathbf{G}_4(x_1, x_2, x_3) = \begin{bmatrix} x_1 & x_2 & x_3 & 0 \\ -x_2^* & x_1^* & 0 & x_3 \\ -x_3^* & 0 & x_1^* & -x_2 \\ 0 & -x_3^* & x_2^* & x_1 \end{bmatrix} \quad (7)$$
  Clearly, this code gives us more flexibility by allowing us to use three different constellations in the same block code as compared to only two constellations as in $\mathbf{G}_2(x_1, x_2)$.

The signal model is given as:

$$\mathbf{Y} = \sqrt{\frac{\rho}{M_t}}\mathbf{C}\mathbf{H} + \mathbf{N} \quad (8)$$

where $\mathbf{C}_{T\times M_t}$ is the transmitted signal matrix and is given as $\mathbf{C} = \sqrt{\frac{T}{K}}\mathbf{G}_{M_t}(x_1, x_2, \ldots, x_K)$; K is the number of different symbols in a codeword. $\mathbf{H}_{M_t\times M_r}$ is the channel coefficient matrix; $\mathbf{Y}_{T\times M_r}$ is the received signal matrix and $\mathbf{N}_{T\times M_r}$ is the noise matrix. The noise samples and the elements of $\mathbf{H}$ are independent samples of a zero-mean complex Gaussian random variable with variance 1. The fading channel is assumed to be quasi-static. The factor $\sqrt{\frac{\rho}{M_t}}$ in (8) is to ensure that ρ is the SNR at each receiver antenna and is independent of M_t. The energy of transmission codeword is normalized to the constrained $E\left\{\|\mathbf{C}\|_F^2\right\} = M_t T$ where $\|\mathbf{C}\|_F^2$ is the Frobenious norm of $\mathbf{C}$. We assume perfect channel state information is known at the receiver, and then the ML decoding is used to minimize the decision metric $\min_{\mathbf{C}} \left\|\mathbf{Y} - \sqrt{\frac{\rho}{M_t}}\mathbf{C}\mathbf{H}\right\|_F^2$ for detecting the transmitted symbols in the codeword, i.e. $x_1, x_2, \ldots, x_K$ independently.

IV. OPTIMAL BANDWIDTH ALLOCATION

We consider the minimization of the expected end-to-end distortion by optimally selecting the quantization parameter (QP) and the GOP size for the source encoder, and the RCPT channel coding rate and the symbol constellation choice for the MIMO transmission. The optimization is constrained on the total available bandwidth. The scalable source encoder produces a layered bitstream where each layer is of different importance, and by protecting these layers unequally using the channel parameters, we can ensure efficient rate allocation between all the layers and then for each of the layers, between the source and the channel coding.

We consider the combined temporal and FGS scalability and define a total of L layers for a GOP. The first $L-2$ layers $(\mu_1, \mu_1, \ldots, \mu_{L-2})$ are the base layers (FGS0) of the frames associated with the lowest to the highest temporal level in decreasing order of importance for video reconstruction. The other two FGS layers (FGS1 and FGS2) of all the frames in a GOP are defined as individual layers (μ_{L-1}, μ_L) of even lesser importance.

The bandwidth allocation problem described above can be formulated as:

$$\min_{\{GOP^*,\ QP^*,\ \mathbf{R}_c^*,\ \mathbf{M}^*\}} E\{D_{s+c}\} \quad s.t. \quad B_{s+c} \le B_{budget}, \quad (9)$$

where B_{s+c} is the transmitted symbol rate, B_{budget} is the total available symbol rate and $E\{D_{s+c}\}$ is the total expected end-to-end distortion which is accurately estimated using the SDDE algorithm as explained in section II. GOP* is the group of picture size for the sequence; $\mathbf{R}_c^* = \{R_{c,\mu_1}, R_{c,\mu_2}, \ldots, R_{c,\mu_L}\}$ specifies the RCPT channel coding rates for each of the layers; $\mathbf{M}^* = \{M_{\mu_1}, M_{\mu_2}, \ldots, M_{\mu_L}\}$ defines the symbol constellations chosen for each of the layers and QP* $= QP_{\mu_{L-2}}$ is the quantization parameter value for the base layer (FGS0) of the highest temporal level. The quantization parameters of all other layers are linearly dependent on $QP_{\mu_{L-2}}$. So all the parameters defined above, GOP*, QP*, $\mathbf{R}_c^*$ and $\mathbf{M}^*$ are optimally selected from a discrete set of admissible values, and are fixed for the complete video sequence. Also according to the O-STBC structure defined in (6) and (7), we restrict the number of optimal

constellation choices for the L layers to a maximum of K different constellations.

It is clear from (5) that the accurate calculation of the decoder distortion depends on the individual probabilities of losing each of the layers for all the frames (P_n, P_{nE1} and P_{nE2}). Let us define the packet error rate for the constant size packets (discussed in section III) as $PER(R_{c,\mu_l}, M_{\mu_l})$, which depends on the channel parameters. Now, the probabilities P_n, P_{nE1} and P_{nE2} are obtained as:

$$P_n = 1 - (1 - PER(R_{c,\mu_l}, M_{\mu_l}))^{\left\lceil \frac{N_{n,\mu_l}}{\gamma} \right\rceil}, \quad l \in \{1, 2, \ldots, L-2\} \tag{10}$$

$$P_{nE1} = 1 - \left(1 - PER\left(R_{c,\mu_{L-1}}, M_{\mu_{L-1}}\right)\right)^{\left\lceil \frac{N_{n,\mu_{L-1}}}{\gamma} \right\rceil} \tag{11}$$

$$P_{nE2} = 1 - (1 - PER(R_{c,\mu_L}, M_{\mu_L}))^{\left\lceil \frac{N_{n,\mu_L}}{\gamma} \right\rceil} \tag{12}$$

where N_{n,μ_l} is the size of FGS0 of the frame n which belongs to the layer μ_l; $N_{n,\mu_{L-1}}$ and N_{n,μ_L} are the size of the layers FGS1 and FGS2 of frame n, respectively.

Next, we define the transmitted symbol rate B_{s+c} as

$$B_{s+c} = \sum_{l=1}^{L} B_{s+c,\mu_l} \tag{13}$$

where B_{s+c,μ_l} is the symbol rate allocated for layer μ_l and is defined by

$$B_{s+c,\mu_l} = \frac{R_{s,\mu_l}}{R_{c,\mu_l} \times \log_2(M_{\mu_l})} \times \frac{T}{K} \tag{14}$$

where R_{s,μ_l} is the source rate for layer μ_l, it is in bits/sec and depends on the quantization parameter value used for that layer (R_{s,μ_L-1} and R_{s,μ_L} depends on the quantization parameter values used for FGS1 and FGS2 respectively).

The problem in (9) is a constrained optimization problem and is solved as an unconstrained one by using the Lagrangian method as in [17], [18].

V. EXPERIMENTAL RESULTS

For the simulations we implemented the SDDE algorithm using the scalable H.264/AVC. The source is taken as 160 frames of two sequences "Foreman" and "Akiyo" at 30 fps with a constant Intra-update (I) at every 16 or 32 frames. After source encoding, the layers are then divided into packets of constant size γ =100 bytes. We consider the admissible set for QP as **QP**={20, 25, 30, 35, 40, 45, 50}, GOP sizes as **GOP**={4, 8, 16} and RCPT coding rates of $\mathbf{R_C} = \{1/3, 1/2\}$ which are obtained by puncturing a mother

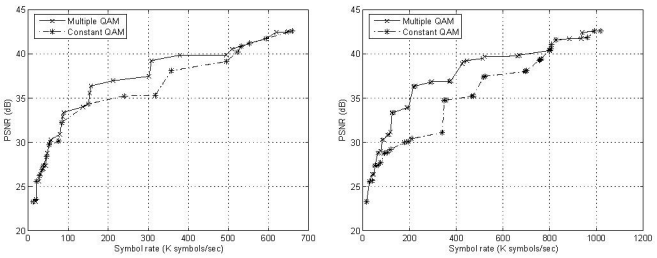
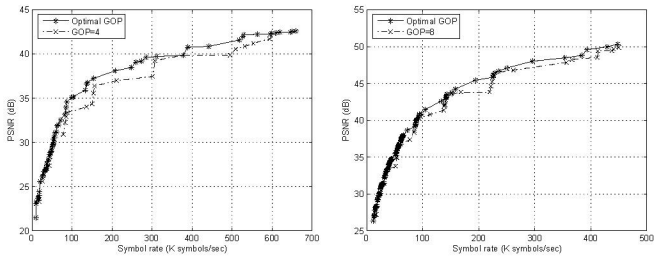


Fig. 3. Comparison of the system performance between multiple QAM versus constant QAM. "Foreman" encoded at GOP = 4 and I=32 for (a) MIMO system 1 and (b) MIMO system 2.

code of rate $\mathbf{R_C} = 1/3$ with constraint length of 3 and a code generator g=$[07;05]_{octal}$. The turbo encoder interleaver size is $L_{inter} = 1600$ bits and is randomly generated. At the RCPT decoder the log-MAP algorithm is used for three iterations. The data are modulated using quadrature amplitude modulation (QAM) with the possible constellations chosen from **M**={4, 8, 16}QAM. At the decoder, scheme 2 is used for error concealment.

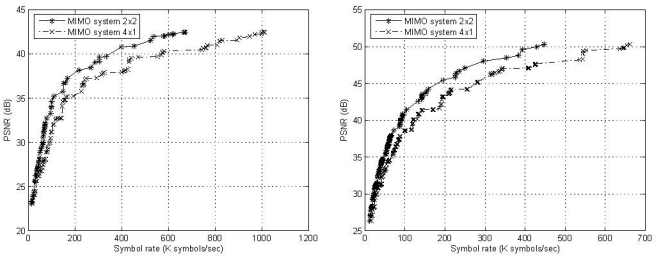
In Figures 3 (a) and (b), we demonstrate the performance of the proposed system for the optimal selection of constellation in comparison with a fixed constellation across all the layers transmitted over the MIMO system. It is clear that using optimal but different modulations across the layers (multiple QAM) outperforms the scenario when the modulation for all the layers is optimally selected but kept fixed (constant QAM) for a wide range of symbol rates. It is also observed that the PSNR improvement gained by using multiple QAM is higher in the MIMO system 2 as compared to MIMO system 1. This can be explained by the fact that MIMO system 2 allows the transmission using three different constellations as compared to a maximum of two different constellations in MIMO system 1. The comparison of the performance of the system for optimal selection of GOP size (for the whole sequence) versus a fixed GOP size for the range of transmission symbol rates is shown in Figures 4 (a) and (b). In these figures, it is evident that using GOP*, QP*, $\mathbf{R_C^*}$ and M* results in better PSNR performance for the reconstructed "Foreman" and "Akiyo" sequences.

In Figures 5(a) and (b), we show the comparison of the two MIMO systems with all the parameters GOP*, QP*, $\mathbf{R_C^*}$ and M* selected after the optimization. Although MIMO system 2 allows more flexibility as compared to MIMO system 1 in selecting the number of constellations, MIMO system 1 still shows better PSNR performance because of the rate of the O-STBC associated with each of the MIMO systems. However,



(a) (b)

Fig. 4. Comparison of the system performance between optimal GOP size versus fixed GOP size for MIMO system 1. (a) "Foreman" encoded at I=32 compared against fixed GOP = 4, and (b) "Akiyo" encoded at I=16 compared against fixed GOP = 8.



(a) (b)

Fig. 5. Comparison of the performance of the two MIMO systems with optimal parameters QP*, GOP*, $\mathbf{R}_{\mathbf{C}}^{*}$ and $\mathbf{M}^{*}$ encoded at I=16 for (a) "Foreman" and (b) "Akiyo" sequences .

MIMO system 2 may be considered more practical to use in scenarios where hardware restrictions limit the number of antennas at the receiver to be only one.

VI. CONCLUSIONS

We proposed a new wireless video transmission system that integrated the latest scalable H.264 coding providing a combined scalability and spatial diversity technique using O-STBC over broadband MIMO systems. We proposed different error concealment schemes to handle packet losses and developed a method for the accurate estimation of the video distortion at the receiver for given channel conditions. The results in comparison with simulated wireless video transmission have shown the accuracy of the distortion estimation algorithm. Using the decoder distortion estimation algorithm, the bandwidth-constrained optimization problem has been solved. We exploited the orthogonal structure of the O-STBC codes used here by allocating different layers over different codeword symbols modulated using different constellations. The results clearly indicate its advantage as compared to using only single constellation for all the layers. Also, it has been shown that optimally selecting GOP size for the whole video sequence for a given bandwidth and channel conditions results in better PSNR performance. The optimal selection for source coding parameters (QP* and GOP*) and channel coding parameters ($\mathbf{R}_{\mathbf{C}}^{*}$ and $\mathbf{M}^{*}$) have been presented for both the MIMO systems for a wide range of symbol rates.

REFERENCES

[1] T. Stockhammer, M. H. Hannuksela, and T. Wiegand, "H.264/AVC in Wireless Environments," *IEEE Trans. CSVT*, vol. 13, no. 7, pp. 657 – 673, Jul. 2003.

[2] S Wenger, "H.264/AVC over IP," *IEEE Trans. CSVT*, vol. 13, no. 7, pp. 645 – 656, Jul. 2003.

[3] H. Schwarz, T. Hinz, H. Kirchhoffer, D. Marpe, and T. Wiegand, "Technical description of the HHI proposal for SVC CE1," ISO/IEC JTC1/SC29/WG11, M11244, October 2004.

[4] H. Schwarz, D. Marpe, T. Schierl, and T. Wiegand, "Combined scalability support for the scalable extension of H. 264/AVC," in *proc. ICME 2005*, Jul. 2005, pp. 446– 449.

[5] J. Reichel, H. Schwarz, and M. Wien, "Scalable working draft - working draft 1," Joint Video Team (JVT), Doc. JVT-N020, Hong Kong, CN, January 2005.

[6] R. Zhang, S.L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE J. Selected Areas in Comm.*, vol. 18, no. 6, pp. 966–976, Jun. 2000.

[7] S. Regunathan, R. Zhang, and K. Rose, "Scalable video coding with robust mode selection," *Signal Proc.: Image Comm.*, vol. 16, pp. 725–732, 2001.

[8] Y. Shen, P. C. Cosman, and L. B. Milstein, "Video coding with fixed-length packetization for a tandem channel," *IEEE Trans. Image Proc.*, vol. 15, no. 2, pp. 273– 288, Feb. 2006.

[9] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Selected Areas in Comm.*, vol. 16, no. 8, pp. 1451–1458, Oct. 1998.

[10] V. Tarokh, N. Seshadri, and A.R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Info. Th.*, vol. 45, no. 5, pp. 1456–1467, Jul. 1999.

[11] J. Song and K. J. R. Liu, "Robust progressive image transmission over OFDM systems using space-time block code," *IEEE Trans. Multimedia*, vol. 4, no. 3, pp. 394–406, Sept. 2002.

[12] S. Zhao, Z. Xiong, X. Wang, and J. Hua, "Progressive video delivery over wideband wireless channels using space-time differentially coded OFDM systems," *IEEE Trans. Mobile Computing*, vol. 5, no. 4, pp. 303– 316, Apr. 2006.

[13] C. Kuo, C. Kim, and C.-C. J. Kuo, "Robust video transmission over wideband wireless channel using space-time coded OFDM systems," in *proc. WCNC2002*, Mar. 2002, vol. 2, pp. 931– 936.

[14] T. Schierl, H. Schwarz, D. Marpe, and T. Wiegand, "Wireless broadcasting using the scalable extension of H. 264/AVC," in *proc ICME 2005*, Jul. 2005, pp. 884– 887.

[15] D. N. Rowitch and L. B. Milstein, "On the performance of hybrid FEC/ARQ systems using rate compatible punctured turbo (RCPT) codes," *IEEE Trans. Comm.*, vol. 48, pp. 948–959, Jun. 2000.

[16] T. V. Ramabadran and S. S. Gaitonde, "A tutorial on CRC computations," *IEEE Micro*, vol. 8, Aug 1998.

[17] L. P. Kondi, F. Ishtiaq, and A. K. Katsaggelos, "Joint source-channel coding for SNR scalable video," *IEEE Trans. Image Proc.*, vol. 11, no. 9, pp. 1043–1054, Sept. 2002.

[18] L. P. Kondi, D. Srinivasan, D. A. Pados, and S.N. Batalama, "Layered video transmission over wireless multirate DS-CDMA links," *IEEE Trans. CSVT*, vol. 15, no. 12, pp. 1629– 1637, Dec. 2005.