Low-Delay Low-Complexity Bandwidth-Constrained Wireless Video Transmission Using SVC Over MIMO Systems

Mohammad K. Jubran, Manu Bansal, and Lisimachos P. Kondi, *Member, IEEE*

***Abstract*—We propose an efficient strategy for the transmission of scalable video over multiple-input multiple-output (MIMO) wireless systems. In this paper, we use the latest scalable H.264 codec (SVC), which provides combined temporal, quality and spatial scalability. At the transmitter, we estimate the decoded video distortion for given channel conditions taking into account the effects of quantization, packet loss and error concealment. The proposed scalable decoder distortion algorithm offers low delay and low complexity. The performance of this method is validated using experimental results. In our 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. The bandwidth constrained allocation problem considered here is simplified and solved for one O-STBC symbol at a time. Furthermore, we take the advantage of the hierarchical structure of SVC to attain the optimal solution for each group of pictures (GOP) of the video sequence. We incorporate the estimated decoder distortion to optimally select the application layer parameter, i.e., quantization parameter (QP), and physical layer parameters, i.e., channel coding rate and modulation type for wireless video transmission.**

***Index Terms*—Distortion estimation, optimal bandwidth allocation, scalable H264, SVC, wireless MIMO systems.**

I. Introduction

THE H.264/AVC standard [1]–[3] and its latest scalable extension [4]–[7], popularly known as SVC, provide superior compression efficiency and have an error-resilient network abstraction layer (NAL) structure for transmission over varied networks. The most important features of SVC are its base layer compatibility with H.264/AVC and the 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. The scalability can be exploited to improve the video transmission over error-prone wireless networks by protecting the different layers with unequal error protection (UEP). This can be achieved by using forward error correction (FEC) combined with an appropriate modulation technique. In many publications it has been shown that under a constrained resource budget, jointly optimizing source and channel coding parameters for scalable video transmission can improve the overall system performance [8]–[12]. FGS is also supported by the MPEG-4 standard. An overview of MPEG-4 FGS and its application to multimedia streaming over IP is presented in [13]. A hybrid temporal/signal-to-noise-ratio (SNR) FGS scheme was proposed in [14]. The unequal packet loss resilience of FGS was explored for the first time in [15]. An FGS scheme with adaptive motion compensation for wireless video was proposed in [16]. FGS with adaptive mode selection and unequal error protection was utilized for wireless video transmission in [17].

In this paper, we consider the efficient transmission of temporal and quality scalable layers over packet-based wireless networks, with optimization of source coding, channel coding and physical layer parameters for each group of pictures (per-GOP basis). A good knowledge of the total end-to-end decoder distortion at the encoder is necessary for such systems. In [18] and [19], a per-pixel based decoder distortion estimation algorithm, the recursive optimal per-pixel estimate (ROPE), was proposed for the non-scalable and SNR scalable H.263+ codec, respectively. Using this algorithm, the first and the second moments of the pixel values, which depend on packet loss probabilities, are recursively obtained to calculate the decoder distortion, which is further used for optimal (source coding) mode selection for a given target rate. In [20], the mean as well as the variance of the end-to-end distortion are considered when allocating limited source and channel resources. In [21], the ROPE algorithm is further modified for different re-synchronization schemes for the transmission of non-scalable H.263 coded video over tandem channels. In this work, we employ our previously proposed scalable decoder distortion estimation (SDDE) algorithm for SVC [22], [23] and propose a new version of the SDDE algorithm with lower delay and complexity. It provides an accurate estimation of the distortion of SVC coded video at the receiver for given channel conditions. The proposed algorithm takes into account loss of both temporal and SNR scalable layers as well as error concealment at the decoder. We compare the performance of the proposed distortion estimation algorithm with both the original SDDE algorithm and the simulated video transmissions over error-prone wireless channels.

Diversity techniques, such as space-time coding (STC) over multiple antenna systems [24]–[26] have been proven to help

Manuscript received May 24, 2007; revised August 09, 2008. Current version published December 10, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hayder Radha.

M. K. Jubran is with the Department of Electrical Engineering, Birzeit University, Birzeit, Palestine (e-mail: mjubran@birzeit.edu).

M. Bansal is with the Intellectual Property Group, Goodwin Procter LLP, Boston, MA 02109 USA (e-mail: mbansal@goodwinprocter.com.

L. P. Kondi is with the Department of Computer Science, University of Ioannina, Ioannina 45110, Greece (e-mail: lkon@cs.uoi.gr).

Digital Object Identifier 10.1109/TMM.2008.2007317

overcome the degradations due to wireless channels (such as fading, the bandlimited nature of the channel, etc.) by providing the receiver with multiple replicas of the transmitted signal over different channels. As one of the STC techniques, orthogonal space-time block codes (O-STBC) were first proposed by Alamouti [24] and later generalized by Tarokh *et al.* [26]. These codes exploit the orthogonality property of the code matrix to achieve the full diversity gain and have the advantage of having a low complexity maximum-likelihood (ML) decoder. We employ these O-STBC codes in the proposed video transmission scheme over the MIMO system and exploit their structure by independently choosing the elements of the codeword from different constellations.

In only a few publications such as [27], [28], wireless video transmission using STC has been studied. In [27], progressive video transmission is proposed over a space-time differentially coded OFDM system with optimal rate and power allocation among multiple layers. In [28], an integrated system of data-partitioned video coding, layered space-time block coding, OFDM modulation and unequal error protection is proposed. It is shown that unequal error protection facilitates the interference cancellation and enhances the quality of reconstructed video, but no optimization for resource allocation is addressed. 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 [29], an approach for using the scalable H.264 with unequal erasure protection (UXP) for temporal scalable over wireless IP networks has been proposed.

Similar to our previous work [22], [23], the proposed system here integrates scalable video coding (temporal and quality scalability) with unequal channel coding using rate-compatible punctured convolutional (RCPC) [30] codes and independent modulation selections for wireless video transmission. The main contributions of this paper are the following. 1) The low-delay and low-complexity version of the SDDE algorithm. This new algorithm prevents the possible drift between encoder and decoder for a more accurate distortion estimation. 2) For the bandwidth constrained optimization problem, we propose specific allocations of the temporal and scalable layers to different O-STBC symbols. The optimization problem is simplified and considered for one O-STBC symbol at a time. The bandwidth allocation problem is addressed by minimizing the expected end-to-end distortion and optimally selecting QP, RCPC rate and the constellation(s) for O-STBC symbols. 3) The optimal parameter selection approach is proposed here on a GOP-by-GOP basis, and is hence more suitable for low-delay applications. Preliminary results of this work were presented in [31].

The rest of the paper is organized as follows. In Section II, we discuss the coding structure of the SVC codec and explain in detail the proposed estimation algorithm for decoder distortion with the error concealment. In Section III, we describe the MIMO system used in this work. In Section IV, we define and address the optimal bandwidth allocation problem. We present the experimental results in Section V. Finally, we present the drawn conclusions in Section VI.

II. Scalable H.264 Codec and Decoder Distortion Estimation

A. Overview of the Scalable Extension of H.264/AVC (SVC)

SVC is based on a hierarchical prediction structure as shown in Fig. 1. A GOP consists of a key picture and all other pictures temporally located between the key picture and the previously encoded key picture. These key pictures are considered as 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). 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. Also, the priority of the base layer (FGS0) of each temporal level decreases from the lowest to the highest temporal level, and each FGS layer for all the frames is considered as a single layer. Further, each layer of each frame is packetized into constant size packets (i.e., $\gamma = 100$ bytes in this work) for transmission. At the receiver, any unrecoverable errors in each packet would result in dropping of that packet and hence would mean the loss of the layer to which the packet belongs. We assume that the base layers of all the key pictures are received error-free. Using the fact that SVC encoding and decoding are done on a GOP basis, it is 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 such that the lost frame is replaced by the nearest available frame in the decreasing as well as increasing sequential order but 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. 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. The video distortion estimation algorithm can be modified to work for any error concealment technique.

B. SDDE TrueRef Derivation

Without loss of generality, in the following derivation of the distortion estimation algorithm, we consider a base layer and two FGS layers. We assume that the frames are lexicographically ordered and the distortion of each macroblock (and hence, each frame) is the summation of the distortion estimated for all the pixels in the macroblock of that frame. 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

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

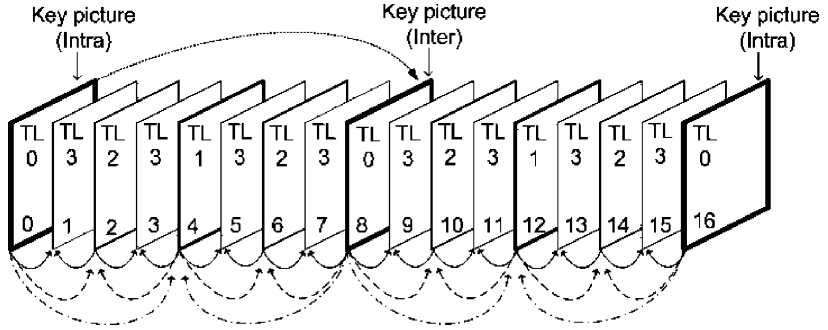


Fig. 1. Hierarchical prediction structure for SVC for GOP size of 8. Two GOPs are shown.

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:

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

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 using 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. We will refer to these frames (u and v) as the "true" reference pictures for frame n. In the decoding process of SVC, 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.

In our previous work [22], [23], [31], the SDDE algorithm considered the frames of $\mathbf{L}_n$ and $\mathbf{R}_n$ and the associated probabilities in estimating the video distortion at the decoder. Here, we propose a low complexity version of the SDDE algorithm by considering the use of only true reference frames for decoding the non-key pictures. This case will be referred to as *SDDE TrueRef*. In this case, if either or both of the true reference frames are not received correctly at the decoder, the non-key picture(s) will be considered erased and will be concealed. The proposed true reference version of SDDE algorithm is aiming to provide better synchronization and prevent possible drift between encoder and decoder. It also reduces the complexity of the SDDE algorithm and hence the processing delay time as now the reference sets are reduced to $\mathbf{L}_n = u$ and $\mathbf{R}_n = v$. For the *SDDE TrueRef* algorithm, the s^{th} moment of the i^{th} pixel of frame n when at least the base layer is received correctly is

$$\begin{aligned} E&\left\{\left(\tilde{f}_n^i(\mathbf{L}_n,\mathbf{R}_n)\right)^s\right\} \\ &= (1-P_{\mathbf{L}_n})(1-P_{\mathbf{R}_n})\,P_{nE1}\left(\hat{f}_{nB\_\mathbf{L}_n\mathbf{R}_n}^i\right)^s \\ &\quad + (1-P_{\mathbf{L}_n})(1-P_{\mathbf{R}_n})\,P_{nE2}\,(1-P_{nE1})\left(\hat{f}_{n(B+E1)\_\mathbf{L}_n\mathbf{R}_n}^i\right)^s \\ &\quad + (1-P_{\mathbf{L}_n})(1-P_{\mathbf{R}_n})(1-P_{nE2})(1-P_{nE1}) \\ &\quad \times\left(\hat{f}_{n(B+E1+E2)\_\mathbf{L}_n\mathbf{R}_n}^i\right)^s \end{aligned} \qquad (3)$$

where, $P_{\mathbf{L}_n}$ and $P_{\mathbf{R}_n}$ are the probabilities of losing the base layer of the reference frames u and v, respectively. Now to get the distortion per-pixel after error concealment, we define a set $\mathbf{Q} = \{f_n, f_{q1}, f_{q2}, f_{q3}, \ldots, f_{GOPend}\}$, 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 $\mathrm{E}\left\{\left(\tilde{f}_n^i\right)^s\right\}$

$$\begin{aligned} \mathrm{E}\left\{\left(\tilde{f}_n^i\right)^s\right\} =& (1-P_n)\,\mathrm{E}\left\{\left(\tilde{f}_n^i(\mathbf{L}_n,\mathbf{R}_n)\right)^s\right\} \\ &+ \left(1-\bar{P}_n\right)(1-P_{q1})\,\mathrm{E}\left\{\left(\tilde{f}_{q1}^i(\mathbf{L}_{q1},\mathbf{R}_{q1})\right)^s\right\} \\ &+ \left(1-\bar{P}_n\bar{P}_{q1}\right)(1-P_{q2})\,\mathrm{E}\left\{\left(\tilde{f}_{q2}^i(\mathbf{L}_{q2},\mathbf{R}_{q2})\right)^s\right\} \\ &+ \cdots + \left(1-\bar{P}_n\prod_{z=1}^{|\mathbf{Q}|-2}\bar{P}_{qz}\right)\mathrm{E}\left\{\left(\tilde{f}_{GOPend}^i\right)^s\right\} \end{aligned} \qquad (4)$$

where $\bar{P}_n = (1-P_n)(1-P_{\mathbf{L}_n})(1-P_{\mathbf{R}_n})$ is the probability of correctly receiving the base layer of frame n and its reference pictures.

The *SDDE TrueRef* algorithm (for each GOP) is summarized as follows.

1) The reconstructed video pixel values ($\hat{f^i_n}$) are obtained using the SVC encoder for a specific QP value and GOP size of interest.

2) For the given values of P_{nE1} and P_{nE2}, the first and second moments of the reconstructed pixel values at the decoder of the key pictures are calculated as in **(2)**.

3) The set of frames $\mathbf{Q}$ is defined based on the error concealment scheme.

4) For the given set of probabilities of losing each of the layer in a GOP, the first and the second moments of the reconstructed pixel values at the decoder of the non-key pictures are calculated based on **(4)** and **(3)**.

5) Having the first and the second moments of the pixel values, the distortion of each pixel [in the mean square sense (MSE)] is calculated using **(1)**. The MSE of all the frames is obtained by averaging the corresponding calculated pixels' distortion.

C. Performance Analysis

The original SDDE and the proposed *SDDE TrueRef* algorithms are implemented by modifying the SVC codec. The *SDDE TrueRef* algorithm performance is evaluated by comparing it with the actual decoder distortion averaged over 200 channel realizations. Its performance is also compared to the original SDDE algorithm. Different video sequences (QCIF format) encoded at 30 fps, GOP size of eight frames and six layers (four temporal levels and two FGS layers) are used in packet-based video transmission simulations. Each of these layers is considered to be affected with different loss rates $P = \{P_{TL0}, P_{TL1}, P_{TL2}, P_{TL3}, P_{E1}, P_{E2}\}$. 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.

In Fig. 2 we compare the performance of the *SDDE TrueRef* algorithms and the actual decoder distortion estimation considering packet loss rates of $P1 = \{0\%, 0\%, 5\%, 5\%, 10\%, 20\%\}$ for both the "Foreman" and "Carphone" sequences considered here. It is evident that the peak SNR (PSNR) values for most of the frames in both sequences closely matches. Similar results are shown in Figs. 3 and 4 considering different packet loss rates $P2 = \{0\%, 10\%, 20\%, 30\%, 50\%, 60\%\}$ and $P3 = \{0\%, 0\%, 10\%, 20\%, 30\%, 40\%\}$, respectively. The average PSNR performance for the actual distortion estimation, the original SDDE algorithm and the *SDDE TrueRef* estimation algorithm is presented in Table I for the "Foreman", "Akiyo" and "Carphone" sequences. As can be observed, the use of the *SDDE TrueRef* algorithm results in comparable average PSNR values as the original SDDE algorithm (with complete reference picture set) for various video sequences.

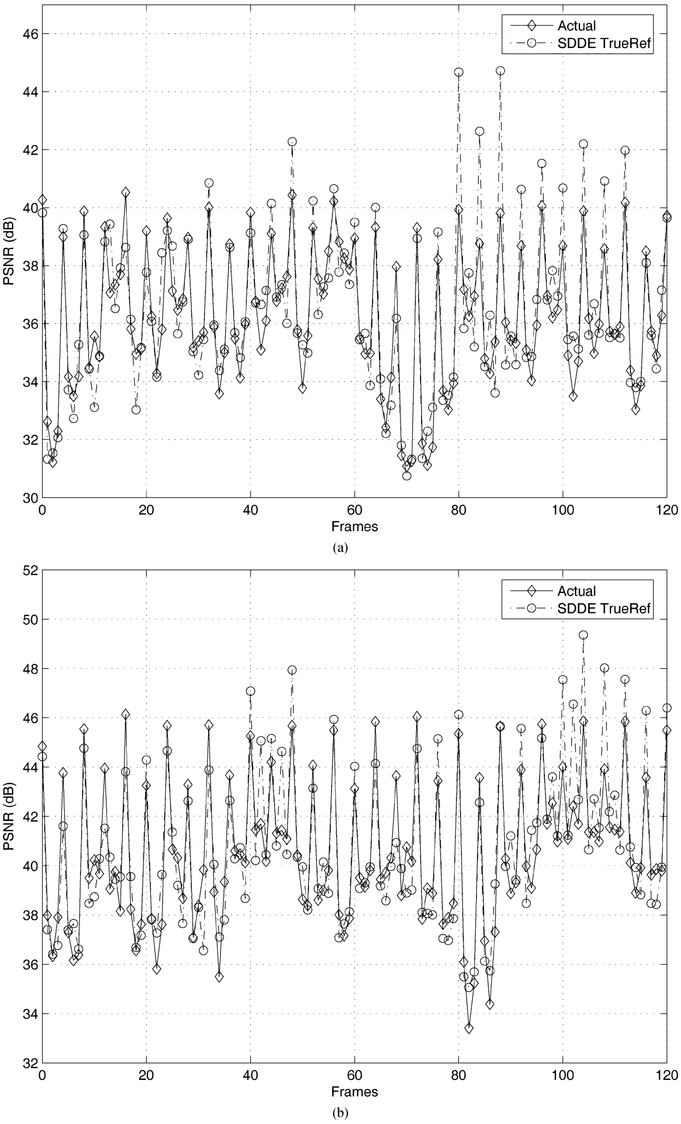
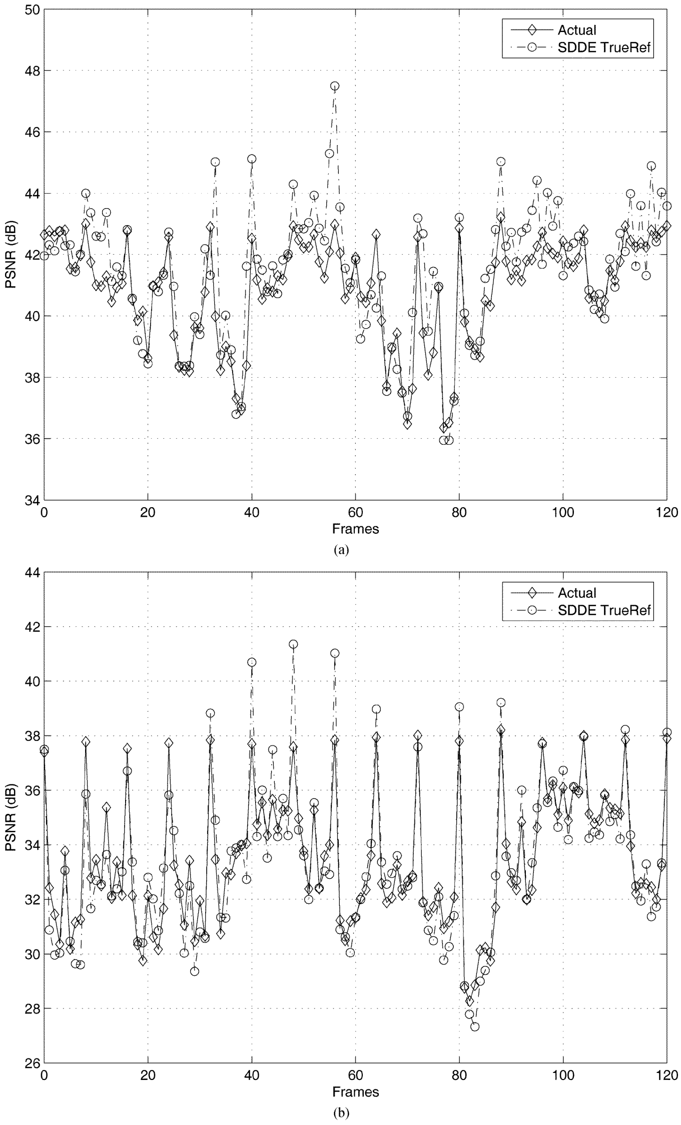


Fig. 2. Comparison between the performance of the actual and the *SDDE TrueRef* algorithm considering packet loss rates of $P1 = \{0\%, 0\%, 5\%, 5\%, 10\%, 20\%\}$. (a) "Foreman" sequence encoded at source rate of 363 kbps. (b) "Carphone" sequence encoded at source rate of 612 kbps.

III. System Description

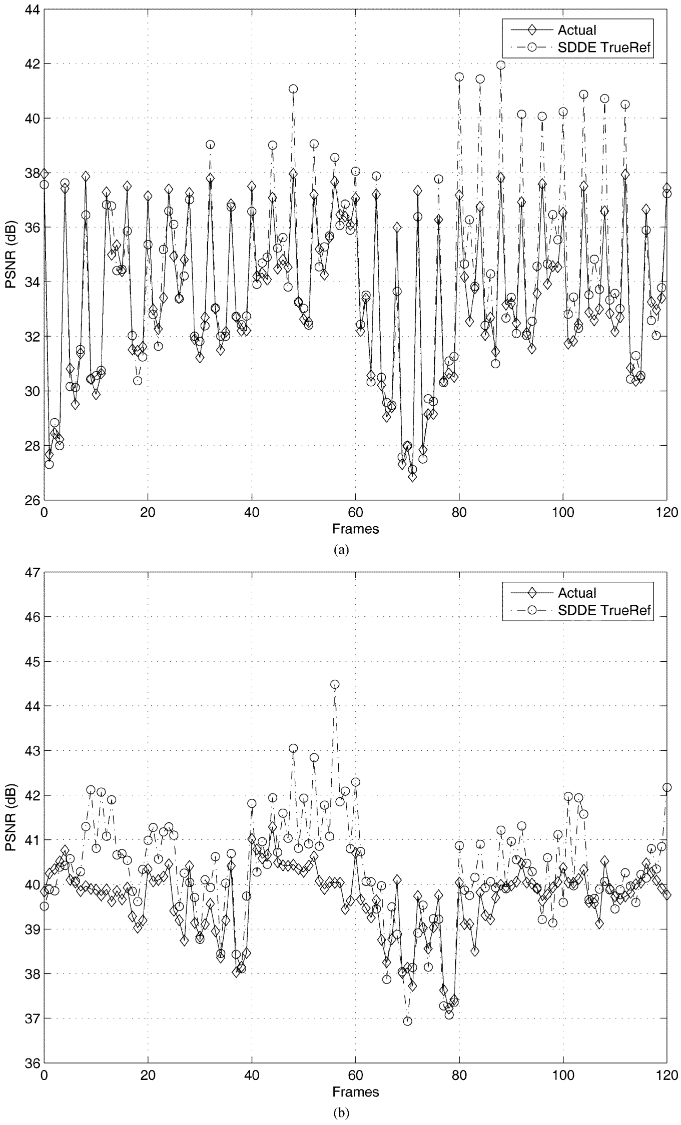
In our packet-based video transmission system, we consider use of channel encoder followed by orthogonal space-time block codes (O-STBC) as shown in Fig. 5. After video encoding, the base and FGS layers of each frame are divided into packets of constant size γ, which are then channel encoded using 16-bit CRC for error detection and rate-compatible punctured convolutional (RCPC) codes for UEP. These channel encoded packets are further encoded using O-STBC for transmission over a MIMO wireless system. A Rayleigh flat-fading channel with AWGN is considered and ML decoding is used to detect the transmitted symbols which are then demodulated and channel decoded for error correction and detection. 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 $M_t = 4$



(a)

(b)

Fig. 3. Comparison between the performance of the actual and the *SDDE TrueRef* algorithm considering packet loss rates of $P2 = \{0\%, 10\%, 20\%, 30\%, 50\%, 60\%\}$, (a) "Akiyo" sequence encoded at source rate of 268 kbps. (b) "Carphone" sequence encoded at source rate of 282 kbps.



(a)

(b)

Fig. 4. Comparison between the performance of the actual and the *SDDE TrueRef* algorithm considering packet loss rate of $P3 = \{0\%, 0\%, 10\%, 20\%, 30\%, 40\%\}$, (a) "Foreman" sequence encoded at source rate of 363 kbps. (b) "Akiyo" sequence encoded at source rate of 157 kbps.

transmit and $M_r = 1$ receive antennas. We used the O-STBC design, $\mathbf{G}_4(x_1, x_2, x_3)$ of rate 3/4 (proposed by Tarokh *et al.* [26]), 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

$$\mathbf{G}_4(x_1, x_2, x_3) = \begin{bmatrix} x_1 & x_2 & x_3 & 0 \\ -x_2^* & x_1^* & 0 & x_3 \\ -x_3^* & 0 & x_1^* & -x_2 \\ 0 & -x_3^* & x_2^* & x_1 \end{bmatrix}. \quad (5)$$

The signal model is given as $\mathbf{Y} = \sqrt{\rho/M_t}\mathbf{CH} + \mathbf{N}$, where $\mathbf{C}_{T\times M_t}$ is the energy-normalized transmitted signal matrix and is given as $\mathbf{C} = \sqrt{T/K}\mathbf{G}_{M_t}(x_1, x_2, \ldots, x_K)$; 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 $\sqrt{\rho/M_t}$ is to ensure that ρ is the SNR at each receiver antenna and is independent of M_t. We assume perfect channel state information is known at the receiver, and the ML decoding is used to detect the transmitted symbols, 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 QP value for video encoding (at the application layer), and the RCPC coding rate and the symbol constellation choice for the MIMO transmission (at the physical layer) on a GOP-by-GOP basis. The optimization is constrained on the total available bandwidth B_{budget}. The available symbol rate is proportional to the available bandwidth. Thus, in the rest of the paper, we will refer to terms

TABLE I
AVERAGE PSNR COMPARISON FOR THE PROPOSED *SDDE TRUE*REF ALGORITHM

	Akiyo 268 kbps	Akiyo 157 kbps	Foreman 829 kbps	Foreman 363 kbps	Carphone 612 kbps	Carphone 282 kbps
Actual P1 (dB)	45.11	41.86	38.32	36.20	40.85	38.54
SDDE P1 (dB)	45.39	42.08	40.18	36.81	42.00	39.16
SDDE TrueRef P1 (dB)	45.81	42.37	39.27	36.46	41.12	38.73
Actual P2 (dB)	40.74	37.87	31.62	30.61	35.32	33.87
SDDE P2 (dB)	41.13	38.21	32.50	31.13	36.33	34.61
SDDE TrueRef P2 (dB)	41.47	38.45	31.69	30.52	35.27	33.91
Actual P3 (dB)	42.84	39.66	35.25	33.46	38.03	36.11
SDDE P3 (dB)	43.36	40.07	37.12	34.36	39.54	37.05
SDDE TrueRef P3 (dB)	43.67	40.23	36.74	34.10	38.78	36.66

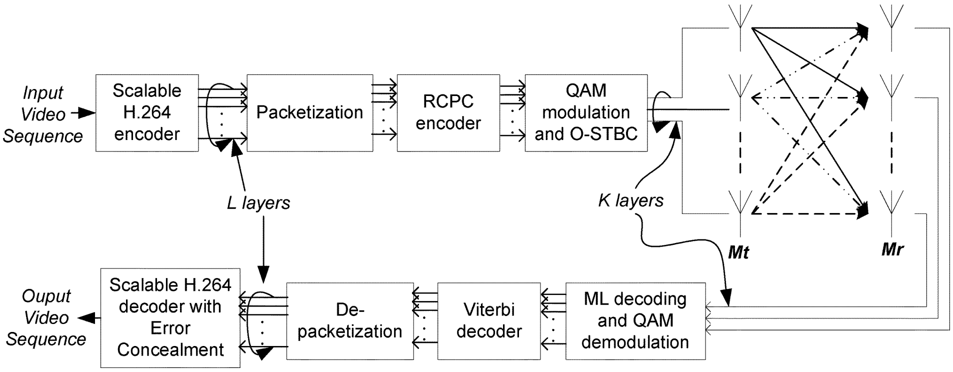


Fig. 5. Block diagram for the scalable H.264 video transmission over MIMO systems.

TABLE II
LAYER ALLOCATION ON O-STBC SYMBOLS

	x_1	x_2	x_3
A_1	$\mu_1, \mu_2, \mu_3, \mu_4$	μ_5	μ_6
A_2	μ_1, μ_2, μ_3	μ_4, μ_5	μ_6
A_3	μ_1, μ_2	μ_3, μ_4, μ_5	μ_6
A_4	μ_1	$\mu_2, \mu_3, \mu_4, \mu_5$	μ_6

"bandwidth" and "symbol rate" interchangeably. We consider the combined temporal and FGS scalability and define a total of L layers for a GOP which are unequally protected by optimally selecting the physical layer parameters. The first $L-2$ layers $(\mu_1, \ldots, \mu_{L-2})$ 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 can be described as

$$\{\mathbf{QP}^*, \mathbf{R}_c^*, \mathbf{M}^*\} = \underset{\{\mathbf{QP}, \mathbf{R}_c, \mathbf{M}\}}{\arg\min} \; E\{D_{s+c}\} \, s.t. B_{s+c} \leq B_{\text{budget}} \tag{6}$$

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 due to source and channel coding which, for given source coding parameters, channel conditions and error concealment, is accurately estimated using the SDDE algorithm as explained in Section II. $\mathbf{QP}$, $\mathbf{R}_c$, and $\mathbf{M}$ are the admissible set of values for QP, RCPC coding rates and symbol constellations, respectively. For all the layers of each GOP, $\mathbf{R}_c^* = \{R_{c,\mu_1}, \ldots, R_{c,\mu_L}\}$; $\mathbf{M}^* = \{M_{\mu_1}, \ldots, M_{\mu_L}\}$ and $\mathbf{QP}^* = \{QP_{\mu_1}, \ldots, QP_{\mu_L}\}$ define the RCPC coding rates, the symbol constellations and QP values, respectively obtained after optimization. The transmitted symbol rate B_{s+c} can be obtained as

$$B_{s+c} = \sum_{l=1}^{L} \frac{R_{s,\mu_l}}{R_{c,\mu_l} \times \log_2(M_{\mu_l})} \times \frac{T}{K} \tag{7}$$

where R_{s,μ_l} is the source coding rate for layer μ_l, it is in bits/s and depends on the quantization parameter value used for that layer; R_{c,μ_l} is the channel coding rate for layer μ_l and is dimensionless; M_{μ_l} is the constellation used by layer μ_l, $\log_2(M_{\mu_l})$ is the bits/symbol and T is the number of time slots required to transmit K symbols in each codeword over the MIMO system.

To solve for the problem defined in (6), we can consider the possible combinations of parameters (from the admissible sets) and obtain the optimal set for all the layers transmitted over different O-STBC symbols or we can take advantage of the independent transmission of each symbol in the O-STBC in (5) by allocating L layers to three different groups corresponding to O-STBC symbols, x_1, x_2 and x_3 as shown in Table II and solve for the bandwidth allocation problem one O-STBC symbol at a time. Here, we will refer to the former as the *no-allocation* case and to the latter as the *allocation* case. It can be seen that for the allocation case, the product space for the possible choices of the parameter is smaller than that of the no-allocation case and hence reduces the complexity of the problem. Table II shows the possible allocations (A_1, A_2, A_3, A_4) considered here for $\text{GOPsize} = 8$ (four temporal and two FGS layers) where each of the six layers is associated with one of the O-STBC symbols. It is necessary to emphasize that here the optimal selection

TABLE III
AVERAGE PSNR VALUES FOR A GOP-BY-GOP ALLOCATION FOR SDDE AND *SDDE TRUE*REF ESTIMATION ALGORITHMS

	Foreman 384 ksps	Foreman 512 ksps	Foreman 768 ksps	Mother-daughter 256 ksps	Mother-daughter 384 ksps
SDDE	37.85	39.30	42.27	40.02	41.63
SDDE TrueRef	37.72	39.16	42.17	39.67	40.87

is done on a GOP-by-GOP basis for each allocation structure and the best allocation is selected after considering the expected distortion (PSNR) criteria for each O-STBC symbol (under the bandwidth constraint). This is defined in more detail as follows.

- Based on **(8)**, the optimal parameter set $\mathbf{X}_1^* = \{\mathbf{QP}_{x_1}^*, \mathbf{R}_{c,x_1}^*, \mathbf{M}_{x_1}^*\}$ for all the layers transmitted over the O-STBC symbol x_1 is obtained by using the admissible set of values of each of the parameter. PSNR_{x_1} and B_{x_1} are the estimated PSNR and the symbol rate allocated for x_1, respectively

$$\mathbf{X}_1^* = \underset{\{\mathbf{QP},\mathbf{R}_c,\mathbf{M}\}}{\arg\max}\ \mathrm{PSNR}_{x_1}\ \ s.t.\ B_{x_1} \leq B_{\mathrm{budget}}. \tag{8}$$

- Given $\mathbf{X}_1^*$, the optimal parameter set $\mathbf{X}_2^*$ is obtained using **(9)**

$$\mathbf{X}_2^* = \underset{\{\mathbf{QP},\mathbf{R}_c,\mathbf{M}\}/\mathbf{X}_1^*}{\arg\max}\ \mathrm{PSNR}_{x_2}\ \ s.t.\ B_{x_2} \leq B_{\mathrm{budget}}. \tag{9}$$

- Finally, having obtained $\mathbf{X}_1^*$ and $\mathbf{X}_2^*$ the optimal set $\mathbf{X}_3^*$ is obtained using **(10)**

$$\mathbf{X}_3^* = \underset{\{\mathbf{QP},\mathbf{R}_c,\mathbf{M}\}/\{\mathbf{X}_1^*,\mathbf{X}_2^*\}}{\arg\max}\ \mathrm{PSNR}_{x_3}\ \ s.t.\ B_{x_3} \leq B_{\mathrm{budget}} \tag{10}$$

where B_{x_a}, $a \in \{1,2,3\}$ is the bandwidth allocated to each O-STBC symbol and is obtained as

$$B_{x_a} = \sum_{\mu_l \in x_a} \frac{R_{s,\mu_l}}{R_{c,\mu_l} \times \log_2(M_{\mu_l})} \times T. \tag{11}$$

PSNR values in (8), (9), and (10) are calculated using the SDDE algorithm. It is clear from Section II-B that the accurate estimation of decoder distortion is dependent upon the probabilities of losing each layer (P_n, P_{nE1} and P_{nE2}). Let us define the packet error rate for the constant size packets ($= \gamma$ bytes) as $PER(R_{c,\mu_l}, M_{\mu_l})$, which depends on the channel parameters. Now, the probabilities P_n, P_{nE1} ($P_n(l = L-1)$) and P_{nE2} ($P_n(l = L)$) are obtained as

$$P_n = 1 – (1 - PER(R_{c,\mu_l}, M_{\mu_l}))^{\lceil N_{n,\mu_l}/\gamma \rceil}, l \in \{1, 2, \ldots, L–2\} \tag{12}$$

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. The problems defined in (6), (8), (9), and (10) are the constrained optimization problems, each of which is solved as an unconstrained one by minimizing the Lagrangian defined as

$$\mathcal{L}(\lambda) = D_{s+c} + \lambda B_{s+c} \tag{13}$$

where λ is the Lagrangian multiplier [8], [10]. The solution to this problem, B_{s+c}^* and hence D_{s+c}^*, is also the solution to the constrained problems if and only if $B_{s+c}^* = B_{\mathrm{budget}}$. In practice, since there is only a finite set of choices for source coding rate, RCPC coding rates and constellation choices, it is not always possible to exactly meet B_{budget}. In this case, the solution is the bandwidth that is closest to B_{budget} while being lower than B_{budget}.

V. EXPERIMENTAL RESULTS

For all the experimental results, "Foreman" and "Mother-daughter" sequences are encoded at 30 fps, GOP $= 8$ and constant Intra-update (I) at every 32 frames. For optimization, we consider QP value in the range of 20 to 50 and RCPC coding rates of $\mathbf{R}_c = 8/N : N \in \{32, 28, 24, 20, 16, 12\}$, which are obtained by puncturing a mother code of rate 8/32 with constraint length of 3 and a code generator $[23; 35; 27; 33]_o$. Quadrature amplitude modulation (QAM) is used with the possible constellations size $\mathbf{M} = \{4, 8, 16\}$.

In Table III, we show the average performance after optimal GOP-by-GOP parameter selection for the system using both SDDE and *SDDE TrueRef* estimation algorithms. The PSNR values shown in Table III are the values obtained by the SDDE and *SDDE TrueRef* distortion estimation algorithms using the optimal parameters selected. It is clear that both algorithms provide comparable estimates of the decoded video distortion for both sequences. Using the optimal parameters selected in the results of Table III, we calculated the expected PSNR at the receiver by simulating the transmission of the encoded video sequence using the packet error rates associated with the optimal channel coding rates and optimal modulation selections. This experiment was repeated 100 times to get the average PSNR. We refer to the PSNR values obtained in this fashion as ”average actual PSNR”. In Table IV, we present the average actual PSNR results for the optimal GOP-by-GOP parameters obtained using the above mentioned SDDE and *SDDE TrueRef* algorithms. Similar to the results from the previous table, the actual PSNR results here are also close for both algorithms. However, the difference in the corresponding values of both the tables is expected as a result of the estimation error as also shown in Section II-C. For all the experimental work presented in the rest of the paper, the original SDDE algorithm will be used.

In Fig. 6, we show the performance of the proposed system for the target bandwidth of 512 kilosymbols per second (ksps) for optimal selection of parameters (on a GOP-by-GOP basis) for

TABLE IV
AVERAGE ACTUAL PSNR RESULTS FOR THE OPTIMAL PARAMETERS SELECTED BASED ON SDDE AND *SDDE TRUE*REF ALGORITHMS

	Foreman 384 ksps	Foreman 512 ksps	Foreman 768 ksps	Mother-daughter 256 ksps	Mother-daughter 384 ksps
SDDE (dB)	39.27	40.40	42.26	40.17	42.49
SDDE TrueRef (dB)	39.04	40.50	42.10	39.37	41.42

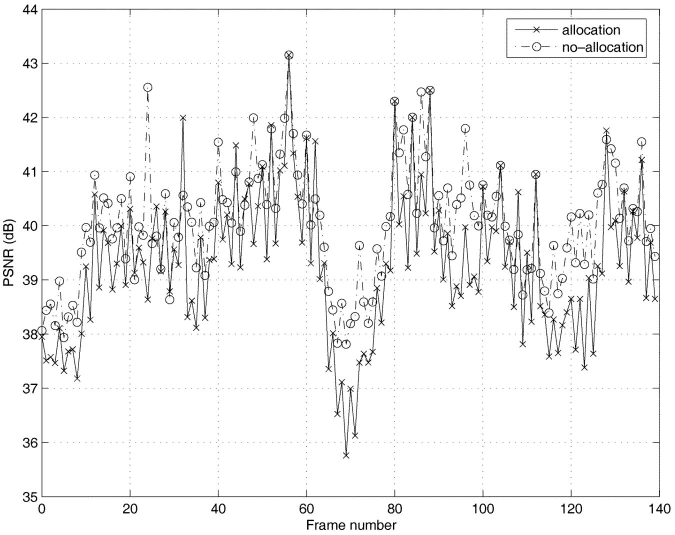


Fig. 6. System performance comparison for target bandwidth of 512 ksps with GOP-by-GOP optimization.

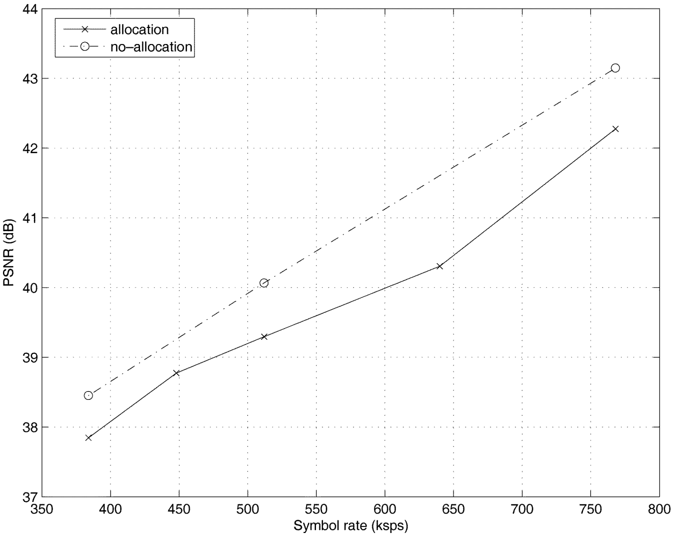


Fig. 7. Average system performance comparison with GOP-by-GOP optimization.

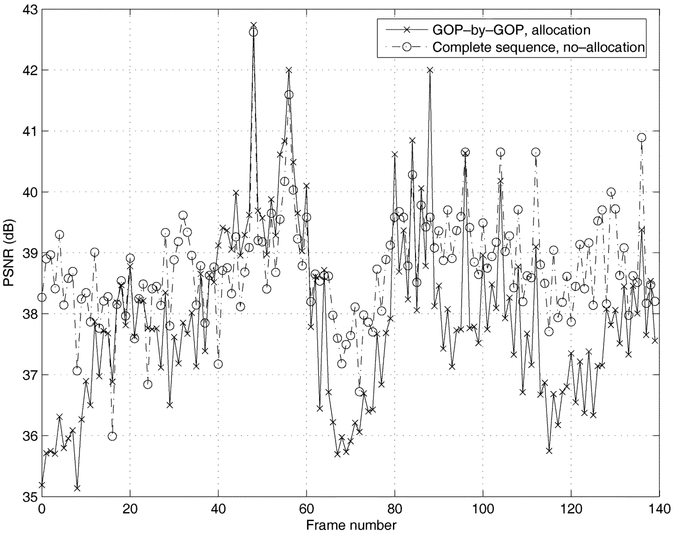


Fig. 8. System performance comparison for target bandwidth of 384 ksps.

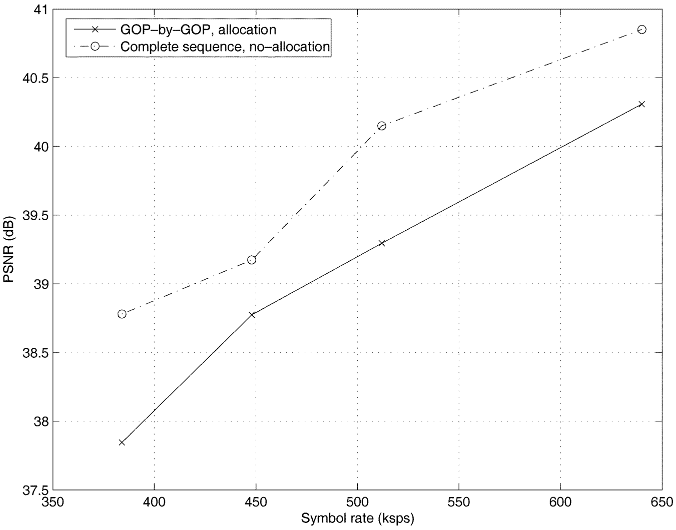


Fig. 9. Average system performance comparison of GOP-by-GOP versus complete sequence optimization.

the allocation and the no-allocation case for the “Foreman” sequence as discussed in Section IV. The comparison in this figure is shown on a per-frame basis and it shows the trade-off between the performance (in the PSNR sense) and the complexity to obtain the optimal solution. As can be seen, the allocation case results in a lower PSNR value (difference of avg. $\mathrm{PSNR} = 0.61\ \mathrm{dB}$), but also has a lower computational complexity than the no-allocation case. It is necessary to mention that the product space for the possible choices of the parameters used in the no-allocation case is much higher than that of the allocation case and hence the latter is of lower complexity. Similar average performance for various target bandwidth values can be seen in Fig. 7. In Figs. 8 and 9 for the “Foreman” sequence, we compare the GOP-by-GOP allocation case with a no-allocation case in which the optimization is carried out on the complete sequence (instead of GOP-by-GOP). The latter case is obviously more complex and incurs an extra delay. Fig. 8 shows the optimal results for a target bandwidth of 384 ksps on a per-frame basis whereas Fig. 9 shows the average comparison for a range of bandwidths. The performance-complexity trade-off is also studied for the scenarios when a complete sequence is considered for optimization problem for both the allocation and the

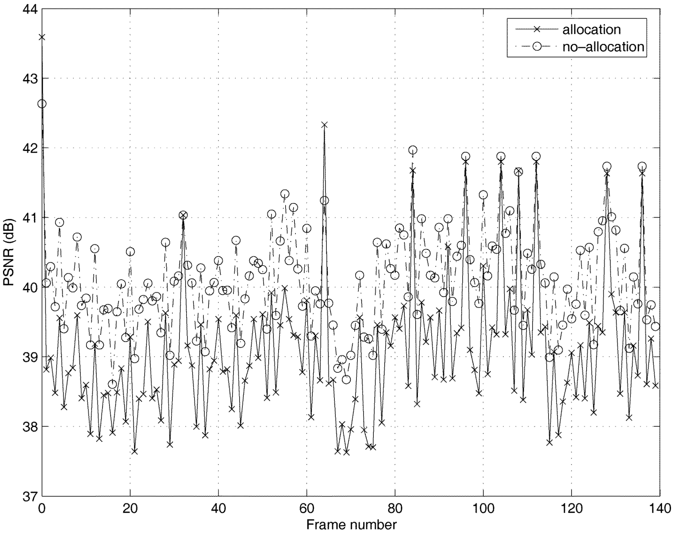


Fig. 10. Performance results for complete sequence optimization with target bandwidth of 512 ksps.

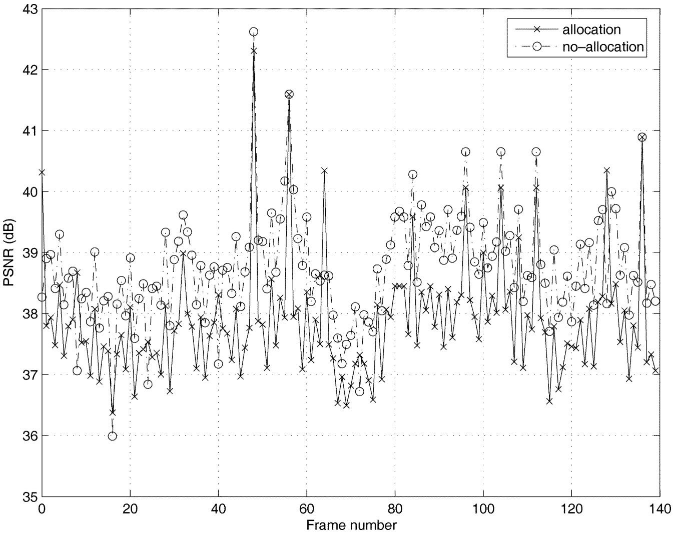


Fig. 11. Performance results for complete sequence optimization with target bandwidth of 384 ksps.

no-allocation cases. The comparison results of this study are shown in Figs. 10 and 11 for the target bandwidths of 512 ksps and 384 ksps, respectively. Fig. 12 shows the optimization results (on a per-frame basis) for the target bandwidths of 384 ksps, 512 ksps, and 640 ksps for the allocation case.

VI. Conclusions

We have proposed a low-delay low-complexity wireless video transmission system that integrated the latest SVC coding with combined scalability and spatial diversity technique using O-STBC over broadband MIMO systems. We have developed a true reference version of the SDDE algorithm that prevents the possible drift between encoder and decoder and is shown to provide comparable performance as the original SDDE algorithm with lower complexity. Using the decoder distortion estimation algorithm, the bandwidth-constrained optimization problem has been simplified and solved. We exploited the orthogonal structure of the O-STBC codes by allocating layers over different codeword symbols modulated using different constellations. The optimization was carried on a GOP-by-GOP basis (suitable for low-delay applications) and the results for different target bandwidth values were presented. The results indicate the advantage of updating the optimal modulation selection every GOP as compared to selecting it optimally but keeping it fixed for the whole sequence.

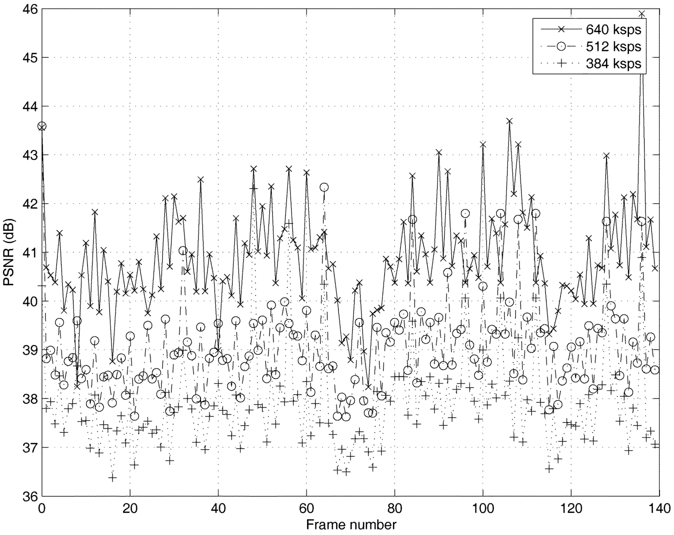


Fig. 12. System performance after optimization for target bandwidths of 640, 512, and 384 ksps.

References

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, “Overview of the H.264/AVC video coding standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 560–576, Jul. 2003.

[2] S. Wenger, “H.264/AVC over IP,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 645–656, Jul. 2003.

[3] T. Stockhammer, M. H. Hannuksela, and T. Wiegand, “H.264/AVC in wireless environments,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 657–673, Jul. 2003.

[4] H. Schwarz, T. Hinz, H. Kirchhoffer, D. Marpe, and T. Wiegand, Technical Description of the HHI Proposal for SVC CE1 2004, ISO/IEC JTC1/SC29/WG11, M11244.

[5] H. Schwarz, D. Marpe, T. Schierl, and T. Wiegand, “Combined scalability support for the scalable extension of H. 264/AVC,” in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME 2005)*, Jul. 2005, pp. 446–449.

[6] J. Reichel, H. Schwarz, and M. Wien, Scalable Working Draft – Working Draft 1 Joint Video Team (JVT), Hong Kong, CN, 2005, Doc. JVT-N020.

[7] H. Schwarz, D. Marpe, and T. Wiegand, “MCTF and scalability extension of H.264/AVC,” in *Proc. PCS 2004*, San Francisco, CA, Dec. 2004.

[8] L. P. Kondi, F. Ishtiaq, and A. K. Katsaggelos, “Joint source-channel coding for motion-compensated DCT-based SNR scalable video,” *IEEE Trans. Image Process.*, vol. 11, no. 9, pp. 1043–1054, Sep. 2002.

[9] F. Zhai, Y. Eisenberg, T. N. Pappas, R. Berry, and A. Katsaggelos, “Rate-distortion optimized hybrid error control for real-time packetized video transmission,” *IEEE Trans. Image Process.*, vol. 15, no. 1, pp. 40–53, Jan. 2006.

[10] L. P. Kondi, D. Srinivasan, D. A. Pados, and S. Batalama, “Layered video transmission over wireless multirate DS-CDMA links,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 12, pp. 1629–1637, Dec. 2005.

[11] Y. S. Chan and J. W. Modestino, “A joint source coding-power control approach for video transmission over CDMA networks,” *IEEE J. Select. Areas Commun.*, vol. 21, no. 12, pp. 1516–1525, Dec. 2003.

[12] D. Srinivasan, L. P. Kondi, and D. A. Pados, "Scalable video transmission over wireless DS-CDMA channels using minimum TSC spreading codes," *IEEE Signal Process. Lett.*, vol. 11, no. 10, pp. 836–840, Oct. 2004.

[13] H. Radha, M. van der Schaar, and Y. Chen, "The MPEG-4 fine-grained scalable video coding method for multimedia streaming over IP," *IEEE Trans. Multimedia*, vol. 3, no. 1, pp. 53–68, Mar. 2001.

[14] M. van der Schaar and H. Radha, "A hybrid temporal-SNR fine-granular scalability for internet video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 318–331, Mar. 2001.

[15] M. van der Schaar and H. Radha, "Unequal packet loss resilience for fine-granular-scalability video," *IEEE Trans. Multimedia*, vol. 3, no. 4, pp. 381–394, Dec. 2001.

[16] M. van der Schaar and H. Radha, "Adaptive motion-compensation fine-granular-scalability (AMC-FGS) for wireless video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 360–371, Jun. 2002.

[17] J. Wu and J. Cai, "Wireless FGS video transmission using adaptive mode selection and unequal error protection," in *Proc. Visual Communications and Image Processing*, San Jose, CA, 2004.

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

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

[20] Y. Eisenberg, F. Zhai, T. N. Pappas, R. Berry, and A. Katsaggelos, "VAPOR: Variance-aware per-pixel optimal resource allocation," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 289–299, Feb. 2006.

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

[22] M. K. Jubran, M. Bansal, R. Grover, and L. P. Kondi, "Optimal bandwidth allocation for scalable H.264 video transmission over MIMO systems," in *Proc. IEEE Military Communications Conf. 2006 (MILCOM 2006)*, Oct. 2006.

[23] M. K. Jubran, M. Bansal, L. P. Kondi, and R. Grover, "Accurate distortion estimation and optimal bandwidth allocation for scalable H.264 video transmission over MIMO systems," *IEEE Trans. Image Process.*, vol. 17, no. 12, Dec. 2008, to be published.

[24] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Select. Areas Commun.*, vol. 16, no. 10, pp. 1451–1458, Oct. 1998.

[25] V. Tarokh, N. Seshadri, and A. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Trans. Inform. Theory*, vol. 44, no. 3, pp. 744–765, Mar. 1998.

[26] V. Tarokh, N. Seshadri, and A. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Inform. Theory*, vol. 45, no. 6, pp. 1456–1467, Jul. 1999.

[27] 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 Comput.*, vol. 5, no. 4, pp. 303–316, Apr. 2006.

[28] C. Kuo, C. Kim, and C.-C. J. Kuo, "Robust video transmission over wideband wireless channel using space-time coded OFDM systems," in *Proc. IEEE Wireless Communications and Networking Conf. (WCNC2002)*, Mar. 2002, vol. 2, pp. 931–936.

[29] T. Schierl, H. Schwarz, D. Marpe, and T. Wiegand, "Wireless broadcasting using the scalable extension of H. 264/AVC," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME 2005)*, Jul. 2005, pp. 884–887.

[30] J. Hagenauer, "Rate-compatible punctured convolutional codes (RCPC codes) and their applications," *IEEE Trans. Commun.*, vol. 36, no. 4, pp. 389–400, Apr. 1988.

[31] M. K. Jubran, M. Bansal, and L. P. Kondi, "Transmission of scalable H.264 coded video over wireless MIMO systems with optimal bandwidth allocation," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP 2007)*, Honolulu, HI, Apr. 2007, pp. I–849–I–852.

Mohammad Jubran received the B.S. degree in electronic engineering from Al-Quds University, Abu-Dies, Palestine, in 2000,and the M.S. and Ph.D. degrees in electrical engineering from the University at Buffalo, The State University of New York, Buffalo, in 2005 and 2007, respectively.

He is currently an Assistant Professor in the Department of Electrical Engineering at Birzeit University, Birzeit, Palestine. His research interests include video compression, joint source/channel coding, multimedia signal processing and transmission, and wireless communications.

Manu Bansal received the B.Tech degree in electronics engineering from YMCA Institute of Engineering, Haryana, India, in 2001 and the M.S. and Ph.D. degrees in electrical engineering from the University at Buffalo, The State University of New York, Buffalo, in 2004 and 2007, respectively.

He is currently a Science Advisor in the Intellectual Property division of Goodwin Procter LLP, Boston, MA. During the summers of 2005–2007, he was an Engineering Intern at Qualcomm, Inc., San Diego, CA, where he worked on the development and implementation of MediaFLO and DVB-H standards of mobile TV technology. His research interests include video processing and communications over wireless networks and joint source-channel coding.

Lisimachos P. Kondi (S'92–M'99) received the Diploma degree in electrical engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1994 and the M.S. and Ph.D. degrees, both in electrical and computer engineering, from Northwestern University, Evanston, IL, in 1996 and 1999, respectively.

During the 1999–2000 academic year, he was a Postdoctoral Research Associate at Northwestern University. He is currently an Assistant Professor in the Department of Computer Science at the University of Ioannina, Ioannina, Greece. He was previously in the faculty of the University at Buffalo, The State University of New York, Buffalo, and has also held summer appointments at the Naval Research Laboratory, Washington, DC, and the Air Force Research Laboratory, Rome, NY. His research interests are in the general areas of signal and image processing and communications, including image and video compression and transmission over wireless channels and the Internet, super-resolution of video sequences, and shape coding.

Dr. Kondi is an Associate Editor for the EURASIP *Journal of Advances in Signal Processing* and an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS.