Unequal Error Protection for Robust Streaming of Scalable Video Over Packet Lossy Networks

Ehsan Maani, *Student Member, IEEE,* and Aggelos K. Katsaggelos, *Fellow, IEEE*

***Abstract*—Efficient bit stream adaptation and resilience to packet losses are two critical requirements in scalable video coding for transmission over packet-lossy networks. Various scalable layers have highly distinct importance, measured by their contribution to the overall video quality. This distinction is especially more significant in the scalable H.264/advanced video coding (AVC) video, due to the employed prediction hierarchy and the drift propagation when quality refinements are missing. Therefore, efficient bit stream adaptation and unequal protection of these layers are of special interest in the scalable H.264/AVC video. This paper proposes an algorithm to accurately estimate the overall distortion of decoder reconstructed frames due to enhancement layer truncation, drift/error propagation, and error concealment in the scalable H.264/AVC video. The method recursively computes the total decoder expected distortion at the picture-level for each layer in the prediction hierarchy. This ensures low computational cost since it bypasses highly complex pixel-level motion compensation operations. Simulation results show an accurate distortion estimation at various channel loss rates. The estimate is further integrated into a cross-layer optimization framework for optimized bit extraction and content-aware channel rate allocation. Experimental results demonstrate that precise distortion estimation enables our proposed transmission system to achieve a significantly higher average video peak signal-to-noise ratio compared to a conventional content independent system.**

***Index Terms*—Channel coding, error correction coding, multimedia communication, video coding, video signal processing.**

I. Introduction

MULTIMEDIA applications involving the transmission of video over communication networks are rapidly increasing in popularity. These applications include but are not limited to multimedia messaging, video telephony, and video conferencing, wireless and wired Internet video streaming, and cable and satellite TV broadcasting. In general, the communication networks supporting these applications are characterized by a wide variability in throughput, delay, and packet loss. Furthermore, a variety of receiving devices with different resources and capabilities are commonly connected to a network. Scalable video coding (SVC) is a highly suitable video transmission and storage system designed to deal with the heterogeneity of the modern communication networks. A video bit stream is called scalable when parts of it can be removed in a way that the resulting substream forms a valid bit stream representing the content of the original with lower resolution and/or quality. Nevertheless, traditionally providing scalability has coincided with significant coding efficiency loss and decoder complexity increase. Primarily due to this reason, the scalable profile of most prior international coding standards such as H.262 MPEG-2 Video, H.263, and MPEG-4 Visual has been rarely used. Designed by taking into account the experience with the past scalable coding tools, the newly developed Scalable Extension of the H.264/advanced video coding (AVC) [1] provides a superb coding efficiency, high bitrate adaptability, and low decoder complexity.

The new SVC standard was approved as Amendment 3 of the AVC standard, with full compatibility of the base layer information so that it can be decoded by existing AVC decoders. The design of the SVC allows for spatial, temporal, and quality scalabilities. The video bit stream generated by the SVC is commonly structured in layers, consisting of a base layer (BL) and one or more enhancement layers (ELs). Each enhancement layer either improves the resolution (spatially or temporally) or the quality of the video sequence. Each layer representing a specific spatial or temporal resolution is identified with a dependence identifier D or temporal identifier T. Moreover, quality refinement layers inside each dependence layer are identified by a quality identifier Q. In some extreme cases, dependence layers may have the same spatial resolution resulting in coarse-grain quality scalability. A detailed description of the SVC can be found in [2]. In this paper, the term SVC is used interchangeably for both the concept of scalable coding in general and for the particular design of the scalable extension of the H.264/AVC standard.

Most modern communications channels (e.g., the Internet or wireless channels) exhibit wide fluctuations in throughput and packet loss rates. Bit stream adaptation in such environments is critical in determining the video quality perceived by the end user. Bit stream adaptation in SVC is attained by deliberately discarding a number of network abstraction layer (NAL) units at the transmitter or in the network before reaching the decoder such that a particular average bit rate and/or resolution is reached. In addition to bit rate adaptation, NAL units may be lost in the channel (due to, for example, excessive delay or buffer overflow) or arrive erroneous at the

Manuscript received January 7, 2009; revised June 11, 2009. First version published November 3, 2009; current version published March 5, 2010. This paper was recommended by Associate Editor J. Ridge.

E. Maani is with the School of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60608 USA (e-mail: ehssan@northwestern.edu).

A. K. Katsaggelos is with the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208 USA (e-mail: aggk@eecs.northwestern.edu).

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

Digital Object Identifier 10.1109/TCSVT.2009.2035846

receiver and therefore have to be discarded by the receiver. A direct approach in dealing with excessive channel losses is to employ error control techniques. However, the optimum video quality is obtained when a circumspect combination of source optimization techniques well-integrated with error control techniques are considered in a cross-layer framework. The benefits of a cross-layer design are considered to be more prominent for scalable video coding since it usually contains various parts with significantly different impact on the quality of the decoded video. This property can be used in conjunction with unequal error protection (UEP) for efficient transmission in communication systems with limited resources and/or relatively high packet loss rates. By using stronger protection for the more important information, error resilience with graceful degradation can be achieved up to a certain degree of transmission errors.

The problem of assigning UEP to scalable video is more complex than that of non-scalable video. The main reason is that scalable video usually consists of multiple scalable layers with different importance in addition to different frame types and temporal dependences. Many researchers have tackled the problem of UEP for scalable video coding by appropriate consideration of the various frame types [3]–[5]. On the other hand, some works have focused on applying UEP to the various quality layers [6]–[8]. For instance, in [7], the impact of applying UEP between base and enhancement layer of fine-granularity-scalability (FGS) coding is studied and the concept of fine-grained loss protection is introduced. Nevertheless, none of the approaches mentioned above jointly considers different frame types (i.e., frame prediction structures) and scalable quality layers. The work presented in [9], on the other hand, jointly considers these two aspects and solves the problem using a genetic algorithm for MPEG-4 scalable video. However, genetic algorithms are considered to be slow and susceptible to *premature convergence* [10].

The aforementioned UEP approaches cannot be directly extended to the SVC coded video, mainly, due to the two new features introduced in the design of the SVC: the hierarchical prediction structure and the concept of key pictures. Unlike prior standards, the prediction structure of the SVC has been designed such that the enhancement layer pictures are typically coded as B-pictures, where the reference pictures are restricted to the temporally preceding and succeeding picture, respectively, with a temporal layer identifier less than the temporal layer identifier of the predicted picture [2]. In addition, the process of motion-compensated prediction (MCP) in SVC, unlike MPEG-4 visual, is designed such that the highest available picture quality is employed for frame prediction in a group of pictures (GOP) except for the *key frames*, i.e., the lowest temporal layer. Therefore, missing quality refinement NAL units of a picture results in propagation of *drift* to all pictures predicted from it. In other words, the distortion of a picture (except for the key frames) depends on the enhancement layers of the pictures from which it has been predicted.

Existing works on robust transmission of SVC using UEP in the literature can be classified into two categories. In the first category, the expected distortion of each frame is estimated and optimized independently by properly allocating source and channel rates [11]. Methodologies developed for joint source channel coding in JPEG2000 such as [12] can also be adapted to be used in SVC under this category. In the second type, the expected distortion of one or more GOPs is estimated and optimized; however, the optimization is carried out using scalable quality layers, i.e., all NAL units within a quality layer are assumed to have the same priority. An example of this approach, presented in [13], uses an approximation model that expresses distortion as a function of bit rate to estimate expected distortion based on the bit rate. [14] employs a more accurate but computationally expensive method to estimate the expected distortion by taking into account the probabilities of losing temporal and/or FGS layers of each frame. Both of these categories ignore the dependences within temporal layers (i.e., the hierarchical prediction structure) and the propagation of drift.

In this paper, we propose a model to accurately and efficiently approximate the per frame expected distortion of the sequence for any subset of the available NAL units and packet loss rates. The proposed model accounts for the hierarchical structure of the SVC, as well as both base and enhancement layer losses. Then, using the proposed distortion model, we address the problem of joint bit extraction and channel rate allocation (UEP) for efficient transmission over packet erasure networks. The rest of this paper is organized as follows. In Section II, we provide an overview of the problem considered and its required components. Subsequently, in Section III we present our distortion and expected distortion calculations. The solution algorithm for both source extraction and joint source-channel coding is then provided in Section IV. Experimental results are shown in Section V and finally conclusion is drawn in Section VI.

II. Problem Formulation

A. Packetization and Channel Coding

Fig. 1 demonstrates the packetization scheme considered in this paper. This scheme has been widely used for providing UEP to layered or progressively coded video, one example is given in [15]. Here, a source packet consists of a SVC NAL unit and portrayed as a row in Fig. 1. Each column, on the other hand, corresponds to a transport layer packet. This figure shows all the source packets included for transmission in one GOP. The source bits and parity bits for the kth source packet are denoted by $R_{s,k}$ and $R_{c,k}$, respectively. The source bits, $R_{s,k}$, are distributed into v_k transport packets and the redundancy bits, $R_{c,k}$, are distributed into the remaining c_k transport packets, as shown in Fig. 1. If a symbol length of m bits is assumed, the length that the kth source packet contributes to each transport packet can be obtained by $l_k = \frac{R_{s,k}}{m v_k}$. Furthermore, channel coding of each source packet is carried out by a Reed–Solomon (RS) code, $RS(N, v_k)$, where N indicates the total number of transport packets in the GOP. Thus, the loss probability of each source packet is given by

$$p_k = 1 - \sum_{i=0}^{t} \binom{i}{N} \epsilon^i (1-\epsilon)^{N-i} \tag{1}$$

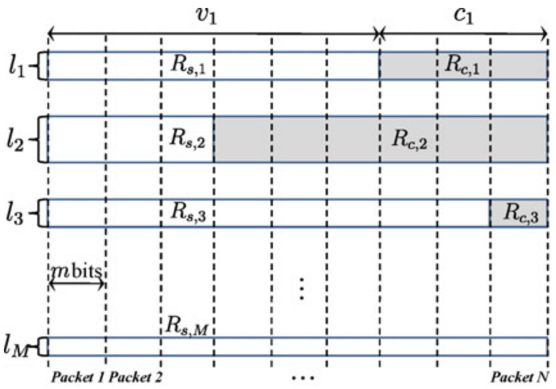


Fig. 1. Structure of channel rate allocation for a GOP.

where ϵ denotes the transport packet loss probability and $t = N - v_k$ is the maximum number of transport packet losses allowed in order to recover the source packet. The channel coding rate assigned to this packet is then v_k/N.

B. System Model

The modes of scalability in SVC are temporal, spatial, and quality scalability. Temporal scalability can be naturally made possible by restricting motion-compensated prediction to reference pictures with a temporal layer identifier that is less than or equal to the temporal layer identifier of the picture to be predicted. In SVC, temporal scalability is provided by the concept of hierarchical B-pictures [16]. Spatial scalability, on the other hand, is achieved by encoding each supported spatial resolution into one layer. In each spatial layer, motion-compensated prediction and intra-prediction are employed similarly to H.264/AVC. The coding efficiency of the SVC is further improved by exploiting additional *inter-layer* prediction mechanisms incorporated into the design of the SVC [2]. Finally, quality scalability is achieved by requantizing the residual signal with a smaller quantization step size relative to that used for the preceding layer. Quality scalability can be seen as a special case of spatial scalability in which the picture sizes for base and enhancement layers are identical. Hence, the same prediction techniques are utilized except for the corresponding upsampling operations. This type of quality scalability is referred to as coarse-grain quality scalable coding (CGS). Since CGS can only provide a few set of decoding points, a variation of the CGS approach, which is referred to as medium-grain quality scalability (MGS), is included in the SVC design to increase the flexibility of bit stream adaptation. MGS coding allows for switching between different MGS layers in any access unit. Furthermore, it is possible to divide the transform coefficient levels to multiple additional MGS layers to achieve finer grain scalability. Each of these MGS layers is identified with a *quality_id* [17]. Fig. 2 portrays the structure of an SVC bit stream with multiple MGS layers.

During transmission, when resources are scarce, a substream of the original SVC bit stream with lower average bit rate is extracted. Commonly, there are a huge number of possibilities (specially for MGS coding) in combining NAL units that result in, approximately, the same bit rate. A very simple method would be to randomly discard NAL units until the desired bit rate is achieved. Nonetheless, the coding efficiency of the resulting bit stream can be significantly compromised if a naive extraction method is used. Consequentially, the concept of quality layers has been incorporated in the architecture of the SVC. To each NAL unit, a *priority identifier* or quality layer related to its contribution to the overall video quality is assigned to be used by the bit stream extractor for efficient adaptation [18]. Optimal bit extraction or assignment of the quality layers in SVC bit streams is a considerably challenging task primarily due to the complications arising from drift propagation. This problem has been considered in [18]–[20]. Here, we consider a joint bit extraction and channel rate allocation to ensure efficient transmission of the SVC streams in lossy environments.

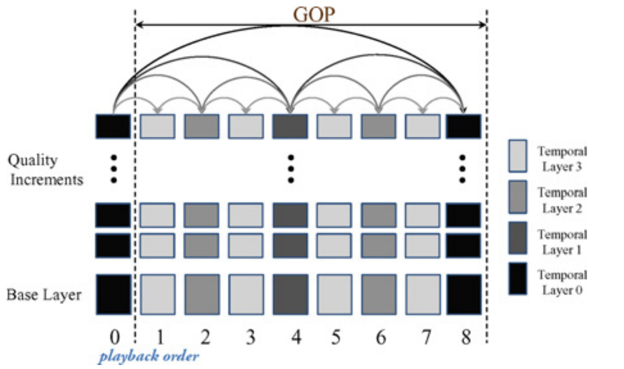


Fig. 2. Structure of a single resolution SVC bit stream.

Let $\pi(n, d, q)$ represent the NAL unit associated with frame n at spatial resolution d and quality level q ($q = 0$ represents the base quality). Then, any "consistent" subset of the NAL units, $\mathcal{P}$, can be uniquely identified by a *selection* map $\phi : \mathbb{Z}^{+2} \to \mathbb{Z}^{+}$ defined by

$$\phi(n, d) = |\mathcal{Q}(n, q)| \tag{2}$$

where $\mathcal{Q}(n, q) := \{q : \pi(n, d, q) \in \mathcal{P}\}$ and the notation $|.|$ represents the cardinality of a set. The term "consistent" here refers to a set whose elements are all decodable by the scalable decoder, i.e., children do not appear in the set without parents. Children here refers to the NAL units that directly depend on others (parents). Note that $\phi(n, d) = 0$ indicates that no NAL unit for frame n at resolution d has been included in the set. When d represents the base resolution, $\phi(n, d) = 0$ means that the base layer of frame n has been skipped and therefore the frames which depend on it through MCP are undecodable. We further define the channel coding function $\psi : \mathbb{Z}^{+3} \to (0, 1]$ such that $\psi(n, d, q)$ denotes the channel rate allocation associated with $\pi(n, d, q)$. Then, the problem of optimal bit extraction and channel rate allocation can be formulated as

$$\begin{aligned}(\boldsymbol{\phi}^*, \boldsymbol{\psi}^*) &= \min_{\boldsymbol{\phi} \in \Phi, \boldsymbol{\psi} \in \Psi} E\{D(\boldsymbol{\phi}, \boldsymbol{\psi}; \epsilon)\} \\ \text{s.t.} \quad & R(\boldsymbol{\phi}, \boldsymbol{\psi}) \leq R_T\end{aligned} \tag{3}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are vector representations of the ϕ and ψ functions, respectively, with element values of $\phi(n, d)$ and

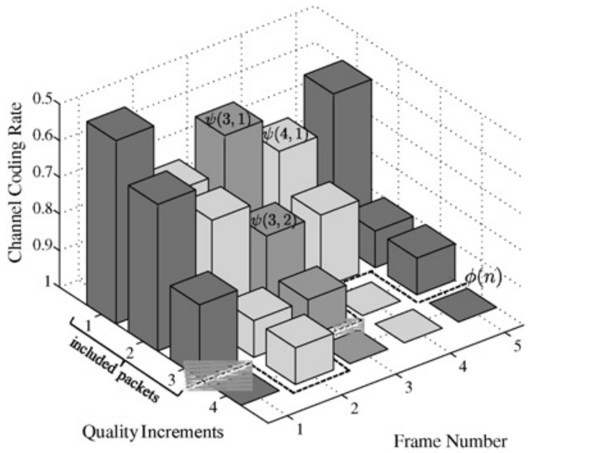


Fig. 3. Example of a selection map and channel rate allocation for a single resolution bit stream.

$\psi(n, d, q)$ for all n, d and q, respectively. Ψ is the set of all possible channel coding rates. Here, due to the nondeterministic nature of channel losses an expected distortion measure is assumed for video quality evaluation. The expected distortion depends on the source packet selection map $\phi(n, d)$ and the associated channel coding rates $\psi(n, d, q)$, as well as the transport packet loss probability ϵ. Further, it should be noted that the variables $\phi(n, d)$ and $\psi(n, d, q)$ are dependent variables since the channel coding rate of a packet is only meaningful if it is included for transmission as indicated by $\phi(n, d)$. In other words, for any possible n and d, $\psi(n, d, q)$ is undefined when $q > \phi(n, d)$. An example of selection and channel coding rate functions for a single resolution bit stream (i.e., d is fixed) is illustrated in Fig. 3.

In principle, a solution to (3) can be found using a non-linear optimization scheme if fast evaluation of the objective functions is possible. However, this problem is characterized by a large number of unknown parameters per sequence or GOP whose optimal values are to be determined. Due to the high dimensionality of the feasible space, a huge number of objective function evaluations are necessary before convergence is reached. Unfortunately, each evaluation of the objective function $E\{D(\boldsymbol{\phi}, \boldsymbol{\psi}; \epsilon)\}$ is highly computationally intense. Various packet loss scenarios with their associated probabilities and reconstructed signal qualities have to be taken into account. Due to the hierarchical prediction structure and existence of drift, evaluation of the video quality for each loss pattern requires decoding of multiple images by performing complex motion compensation operations. Consequentially, the computational burden of this optimization is considered to be far away from being manageable. As a solution, in the next section we propose a computationally efficient and yet accurate model that provides an estimate of the sequence distortion for any selection map $\boldsymbol{\phi}$ and channel rate allocation function $\boldsymbol{\psi}$.

III. Expected Distortion Calculations

As discussed in Section II-B, fast evaluation of the sequence expected distortion plays an essential role in solving the optimization problem of (3) and thus constitutes the main contribution of our paper. In this section, we introduce an approximation method for the computation of this distortion. For this purpose, we consider a single-resolution SVC stream in this paper. Nonetheless, our calculations can be directly applied to the more general multiresolution case if we assume that all quality NAL units associated with lower resolution spatial layers are included before the base quality of a higher resolution. This constraint reduces the degrees of freedom associated with the selection and channel rate allocation functions by one. Hence, they can be denoted by $\phi(n)$ and $\psi(n, q)$, respectively. Regardless of the number of spatial layers in the SVC bitstream, a target resolution has to be specified to evaluate the quality of the reconstructed sequence. The quality increments from spatial layers lower than the target resolution need to be up-sampled to the target resolution to evaluate their impact on the signal quality. The video quality is measured using the mean square error metric with respect to the fully reconstructed signal. The reason for this is that the considered system is a transmission system often implemented separately from the encoder and thus has no access to the original uncompressed signal.

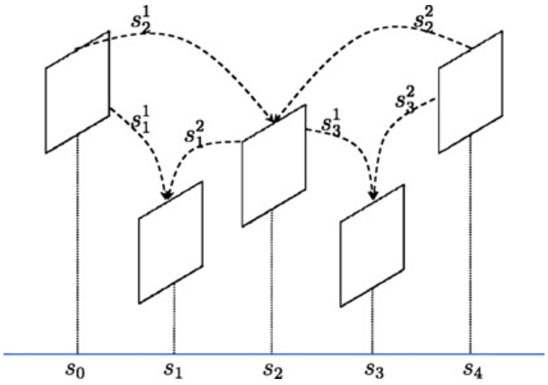


Fig. 4. Hierarchical prediction structure in a GOP of size 4.

For applications in which transmission over a packet lossy network is required, the expected distortion has to be considered to evaluate the video quality at the encoder. Our expected distortion model assumes knowledge of channel state information and the particular error concealment method employed by the decoder. In this paper, a simple and popular concealment strategy is employed: the lost picture is replaced by the nearest temporal neighboring picture. The expected distortion of a GOP is calculated based on the selection function $\phi(n)$ of the GOP. As mentioned in Section II-B for the general case, $\phi(n)$ specifies the number of quality increments to be sent per frame n. We consider a generic case where a packet loss probability of p_n^q is assigned to the qth quality increment packet of frame n, i.e., $\pi(n, q)$. Recall that p_n^q is dependent on the transport packet loss probability and the specific channel coding rate $\psi(n, q)$. Additionally, let the set $\mathcal{S} = \{s_0, s_1, ..., s_N\}$ represent the N pictures in the GOP plus the key picture of the preceding GOP denoted by s_0 as portrayed in Fig. 4 (for $N = 4$). We further define a function $g : \mathcal{S} \rightarrow \mathbb{Z}$ such that $g(x)$ indicates the display order frame number of any $x \in \mathcal{S}$. Note that in our notation the nth frame (in display order) is denoted as n and s_n, interchangeably.

Let $\tilde{D}_n$ denote the distortion of frame n after decoding as seen by the encoder, i.e., $\tilde{D}_n$ represents a random variable whose sample space is defined by the set of all possible distortions of frame n at the decoder. Then, assuming that a total number of Q quality levels exist per frame, the conditional expected frame distortion $E\{\tilde{D}_n|BL\}$ given that the base layer is received intact is obtained by

$$\begin{aligned} E\{\tilde{D}_n|BL\} = \sum_{q=1}^{\phi(n)} p_n^q D_n(q-1) \prod_{i=0}^{q-1}(1-p_n^i) \\ + D_n(\phi(n)) \prod_{i=0}^{\phi(n)}(1-p_n^i) \end{aligned} \tag{4}$$

where $D_n(q)$ is the total distortion of frame n reconstructed by inclusion of $q > 0$ quality increments. The first term in (4) accounts for cases in which, all $(q-1)$ quality segments have been successfully received but the qth segment is lost, therefore, the reconstructed image quality is $D_n(q-1)$. The second term, on the other hand, accounts for the case where all quality increments in the current frame sent by the transmitter [given by $\phi(n)$] are received.

Due to the hierarchical coding structure of the SVC, decoding of the base layer of a frame not only requires the base layer of that frame but also the base layers of all preceding frames in the hierarchy which were used for the prediction of the current frame. For instance, decoding any of the frames in the GOP requires that the key picture of the preceding GOP, s_0, be available at the decoder. We define a relation $\preceq$ on the set $\mathcal{S}$ such that if $x, y \in \mathcal{S}$ and $x \preceq y$ then x depends on y via motion-compensated prediction; x is referred to as *child* of y if it is directly predicted from y. For each frame $s_n \in \mathcal{S}$, a set Δ_n can be formed consisting of all reference pictures in $\mathcal{S}$ that the decoder requires in order to decode a base quality of the frame. This set is also referred to as the ancestor frames set. It can be verified that the set Δ_n plus the relation $\preceq$ on the set form a *well-ordered* set since all four properties, i.e., reflexivity, antisymmetry, transitivity, and comparability (trichotomy law) hold. Note that because all frames in the GOP depend on the key picture of the preceding GOP and no frame in Δ_n depends on frame s_n, for all $n \leq N$ we have

$$s_0 \succeq x,\ s_n \preceq x,\ \forall x \in \Delta_n(x). \tag{5}$$

In the case that the base layer of a frame $x \in \Delta_n$ is lost, the decoder is unable to decode frame n and therefore has to perform concealment from the closest available neighboring frame in display order. If we denote this frame by k, then the distortion of frame n after concealment can be represented by $D_{n,k}^{con}$. Consequently, the expected distortion of frame n is computed according to

$$\begin{aligned} E\{\tilde{D}_n\} = \sum_{i\in\Delta_n} p_i^0\, D_{n,k}^{con} \prod_{\substack{j\in\Delta_n \\ j\prec i}} (1-p_j^0) \\ + E\{\tilde{D}_n|BL\} \prod_{j\in\Delta_n} (1-p_j^0) \end{aligned} \tag{6}$$

where k represents the concealing frame, s_k, specified as the nearest available temporal neighbor of i, i.e.,

$$s_k = \arg\min_{\substack{x\in\Delta_n \\ i\prec x}} |g(x) - g(s_i)|. \tag{7}$$

Here, $g(x)$ indicates the display order frame number as defined before. The first term in (6) deals with situations in which the base layer of a predecessor frame i is lost (with probability p_i^0) and thus frame n has to be concealed using a decodable temporal neighbor while the second term indicates the case in which all base layers are received.

From (4) and (6), it is apparent that the calculation of the expected distortion $E\{\tilde{D}_n\}$ requires computation of $D_n(q)$ for all $q < Q$ and $D_{n,k}^{con}$ for various concealment options. $D_n(q)$ refers to the total distortion of frame n if $q > 0$ quality increments are received (it is assumed that the base layer has been received). Note that even though $D_n(q)$ refers to the case where q quality increments have been successfully received for s_n, it still represents a non-deterministic variable since the number of quality increments received for the ancestor frames of s_n (Δ_n) is unknown. For situations in which the base quality of the nth frame cannot be reconstructed the decoder performs error concealment. The frame distortion in this case is given by $D_{n,k}^{con}$. Below, we discuss the computations of $D_n(q)$ and $D_{n,k}^{con}$ in detail.

A. Frames With Decodable Base Layer

Since for MGS coding of SVC, motion compensated prediction is conducted using the highest available quality of the reference pictures (except for the key frames), propagation of drift has to be taken into account whenever a refinement packet is missing. Let $\boldsymbol{f}_n^d$ and $\boldsymbol{f}_n$ denote a vector representation of the reconstructed nth frame using all of its quality increments in the presence and absence of drift, respectively. Note that although all quality increments of frame n are included for the reconstruction of both $\boldsymbol{f}_n$ and $\boldsymbol{f}_n^d$, in general $\boldsymbol{f}_n^d \neq \boldsymbol{f}_n$ since some quality increment of the parent frames may be missing in the reconstruction of $\boldsymbol{f}_n^d$. Furthermore, missing quality increments of frame n introduce additional degradation. Let $\boldsymbol{e}_n(q)$ represent the error vector resulting from the inclusion of $q \leq Q$ quality increments for the nth frame in the absence of drift. It should be noted that this error vanishes when all refinements of the frame are added, i.e., $\boldsymbol{e}_n(Q) = \boldsymbol{0}$. We refer to this error as the EL clipping error. The total distortion of frame n due to drift and EL clipping (i.e., D_n) with respect to $\boldsymbol{f}_n$ is obtained according to

$$\begin{aligned} D_n(q) &= ||\boldsymbol{f}_n - \boldsymbol{f}_n^d + \boldsymbol{e}_n(q)||^2 \\ &= D_n^d + D_n^e(q) + 2(\boldsymbol{f}_n - \boldsymbol{f}_n^d)^T \boldsymbol{e}_n(q) \end{aligned} \tag{8}$$

where D_n^d and $D_n^e(q)$ represent, respectively, the distortion, i.e., sum of squared errors, due to drift and EL clipping (associated with the inclusion of q quality increments). The symbol $||.||$ here represents the l_2-norm. Since the Cauchy–Schwartz inequality provides an upper bound to (8) we can approximate the total distortion D_n as

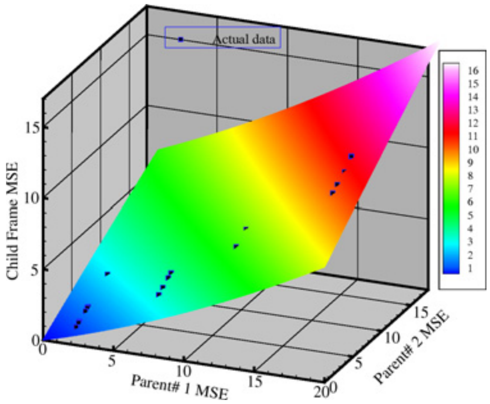


Fig. 5. Example of a parent–child distortion relationship with a quadratic surface fit.

$$
\begin{aligned}
D_n(q) &\approx D_n^d + D_n^e(q) + 2\kappa\sqrt{D_n^d}\sqrt{D_n^e(q)} \\
&\le D_n^d + D_n^e(q) + 2\sqrt{D_n^d}\sqrt{D_n^e(q)}
\end{aligned}
\tag{9}
$$

where κ is a constant in the range $0 \le \kappa \le 1$ obtained experimentally from test sequences. In consequence, to calculate the total distortion, we only need the drift and EL clipping distortions, D_n^d and $D_n^e(q)$, respectively. Fortunately, the error due to EL clipping, $D_n^e(q)$, can be easily computed by inverse transforming the de-quantized coefficients read from the bit stream. The drift distortions, on the other hand, depend on the computationally intensive motion compensation operations and propagate from a picture to its descendants.

Similarly, to the definition of the ancestor set Δ_n we can define a parent set Λ_n to include the two parents of s_n referred to as s_n^1 and s_n^2. For instance, the parent set for frame s_2 in Fig. 4 equals $\Lambda_2 = \{s_0, s_4\}$. Further, let D_i represent the total distortion of a parent frame of s_n, where, $i \in \Lambda_n$. Then, we can assume that the drift distortion inherited by the child frame, denoted by D_n^d, is a function of parent distortions, i.e., $D_n^d = F(D_{s_n^1}, D_{s_n^2})$. Therefore, an approximation to D_d^n can be obtained by a second order Taylor expansion of the function F around zero

$$
D_n^d \approx \gamma + \sum_{i\in\Lambda_n} \alpha_i D_i + \sum_{i\in\Lambda_n}\sum_{j\in\Lambda_n} \beta_{ij} D_i D_j. \tag{10}
$$

Here the coefficients α_i and β_{ij} are first and second order partial derivatives of F at zero and are obtained by fitting a 2-D quadratic surface to the data points acquired by the decodings of the sequence/GOP with a limited number of different reconstructed qualities. The constant term $\gamma = 0$ since there is no drift distortion when both reference frames are fully reconstructed, i.e., $D_i = 0$, $i = 1, 2$. Note that technically, $F(D_{s_n^1}, D_{s_n^2})$ is not a function since the mapping $\{D_{s_n^1}, D_{s_n^2}\} \rightarrow D_n^d$ is not a unique mapping because distortions may be due to various error distributions. Therefore, (10) can only be justified as an approximation. It should be noted that the coefficients α_i and β_{ij} are computed per frame and are specific to a single SVC bit stream reflecting the characteristics of that bit stream. Fig. 5 demonstrates an example of this parent–child distortion relation for a frame of the *Foreman* sequence. The coefficients of this equation for all frames except the key frames can be obtained by several decodings of different substreams extracted from the global SVC bit stream. Nevertheless, different methods for choosing the data points may exist. For instance, a simple method to acquire these data points is described in [20].

Once the coefficients α_i and β_{ij} are computed for each frame (except for the key frames), the drift distortion of the child frame D_n^d can be estimated depending on the distortion of the parent frames according to (10). Note that the actual distortion of the parent frames is unknown to the transmitter; thus, an expected value of the parent distortion has to be used. Since a decodable base layer is assumed in this section, the drift distortions are obtained as

$$
\begin{aligned}
D_n^d \approx &\sum_{i\in\Lambda_n} \alpha_i E\{\tilde{D}_i|BL\} \\
&+ \sum_{i\in\Lambda_n}\sum_{j\in\Lambda_n} \beta_{ij} E\{\tilde{D}_i|BL\}E\{\tilde{D}_j|BL\}.
\end{aligned}
\tag{11}
$$

The total distortion $D_n(q)$ is then computed according to (9). Note that since the drift distortions depend on the qualities of the parent frames, for each GOP the expected distortion computation has to start from the highest level in the prediction hierarchy, i.e., the key frame, for which $D_n^d = 0$. Once the total distortion of the key frame is attained, its expected distortion given the base layer $E\{\tilde{D}_n|BL\}$ can be calculated as described by (4). This value is then used to find the drift distortion of the child frame utilizing the above equation. This drift distortion then yields to the computations of $D_n(q)$ and $E\{\tilde{D}_n|BL\}$ for the child frame according to (9) and (4), respectively. This process continues for the children of the child frame until the conditional expected distortions $E\{\tilde{D}_n|BL\}$ are computed for the entire GOP.

B. Frames With Missing Base Layer

The base quality NAL unit may be skipped at the transmitted or be damaged or lost in the channel and therefore become unavailable to the decoder. In this scenario, all descendants of the frame to which the NAL unit belongs to are also discarded by the decoder and an error concealment technique is utilized. To be able to determine the impact of a frame loss on the overall quality of the video sequence, the distortion of the lost frame after concealment needs to be computed.

As before, let $D_{n,i}^{con}$ denote the distortion of a frame n concealed using frame i with a total distortion of D_i. From our experiments, we observed that $D_{n,i}^{con}$ does not vary noticeably with respect to D_i, therefore we use a first order expansion to approximate $D_{n,i}^{con}$, i.e.,

$$
D_{n,i}^{con} \approx \mu_i + \nu_i D_i \tag{12}
$$

where μ_i and ν_i are constant coefficients calculated for each frame with all concealment options (different i's). In a high activity video sequence, due to the content mismatch between frame n and i we have $\mu_i \gg \nu_i$. On the other hand, for low activity sequence, we expect $\mu_i \approx 0$ and $\nu_i \approx 1$. For each frame, there are usually multiple concealment options, as an example, in Fig. 4, the concealment options for frame s_3, in the preferred order, are $\{s_2, s_4, s_0\}$. The coefficients in (12) are obtained by conducting a linear regression analysis on the actual data points. Note that these data points are acquired by performing error concealment on frames reconstructed from decodings explained in Section III-A. Hence, no extra decoding is required for this process. Based on the above discussion, the distortion after concealment $D_{n,k}^{con}$ in (6) is computed by

$$D_{n,k}^{con} \approx \mu_k + \nu_k E\{\tilde{D}_k|BL\}. \tag{13}$$

Finally, with the calculation of $D_{n,k}^{con}$, the overall expected distortion of the entire GOP can be estimated using (6). Recall that the conditional expected distortions $E\{\tilde{D}_k|BL\}$ are known by this time as discussed in Section III-A. In order to evaluate the accuracy of the proposed expected distortion model, we compared the calculated expected distortion to an average of the decoded distortion for various loss patterns. Fig. 6 shows an example of this comparison for the *Foreman* common intermediate format (CIF) sequence. A random selection map is first generated for the sequence, then, according to the selection map packets are either discarded or transmitted through a channel with pre-defined loss probabilities (no channel coding was considered). The solid line shows the average per-frame distortions obtained by considering 500 channel realizations, while the dashed line represents the estimated distortions computed using the proposed method. Moreover, the grey area indicates the standard deviation of the reconstructed signal quality overall channel realizations.

IV. Solution Algorithm

The distortion model proposed, in this paper, allows for accurate and fast computation of the expected distortion of the SVC bit streams transmitted over a generic packet lossy network. In this section, utilizing this distortion model, we develop an algorithm to perform joint bit extraction and channel rate allocation for robust delivery of SVC streams. Note that according to (4) and (6), the expected distortion of the video sequence directly depends on the source mapping function $\phi(n)$. Its dependence on the channel coding rates, on the other hand, is implicit in those equations. The source packet loss probabilities, p_n^q's, used for the computation of the expected distortion depend on the channel conditions, as well as the particular channel coding and rate employed as shown in (1).

The optimization can be performed over an arbitrary number of GOPs, denoted by M. Note that increase in the size of the optimization window, M, may result in a greater performance gain but at a price of higher computational complexity. The source mapping function $\phi(n)$ initially only includes the base layer of the key pictures with an initial channel coding rate of 1. Then, at each time step, a decision is made whether to add a new packet to the transmission queue or increase the forward error correction protection of an existing packet. Among all already included packets in the transmission queue, we identify a $\pi(n^*, q^*)$ such that an increase in its channel protection results in the highest expected distortion gradient, δED^*. Thus, we have

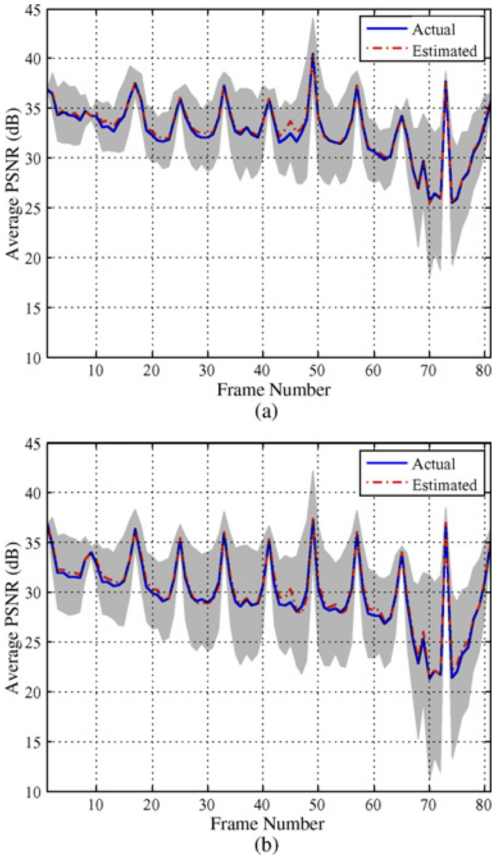


Fig. 6. Actual versus estimated frame distortions for various loss probabilities; grey area denotes the standard deviation of the actual distortions. (a) $p = 5\%$. (b) $p = 15\%$.

$$\delta ED^* = \max_{n} \max_{q<\phi(n)} \left|\frac{\partial ED(\boldsymbol{\phi}, \boldsymbol{\psi})/\partial\psi(n,q)}{\partial R_t(\boldsymbol{\phi}, \boldsymbol{\psi})/\partial\psi(n,q)}\right| \tag{14}$$

where ED and R_t represent the expected distortion and the total rate associated with the current $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$. Here, the constraint $q < \phi(n)$ ensures that the packet has already been included in the selection map at a preceding time step. Likewise, among the candidate packets for inclusion, let $\pi(n^\dagger, \phi(n^\dagger))$ denote the one with highest expected distortion gradient, $\delta ED^\dagger$, i.e.,

$$\delta ED^\dagger = \max_{n} \max_{\psi(n,q)\in\Psi} \left|\frac{\partial^2 ED(\boldsymbol{\phi}, \boldsymbol{\psi})/\partial\phi(n)\partial\psi(n,q)}{\partial^2 R_t(\boldsymbol{\phi}, \boldsymbol{\psi})/\partial\phi(n)\partial\psi(n,q)}\right| \tag{15}$$

where $q = \phi(n)$. In cases for which $\delta ED^* > \delta ED^\dagger$, the channel protection rate of the already included packet $\pi(n^*, q^*)$ is incremented to the next level by padding additional parity bits. Conversely, when $\delta ED^* < \delta ED^\dagger$, the source packet $\pi(n^\dagger, \phi(n^\dagger))$ is included in the transmission queue with a channel coding rate $\psi(n^\dagger, \phi(n^\dagger))$ obtained from (15). Note that in both scenarios, the corresponding functions $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are updated according to the changes made to the transmission queue. This process is continued until the bit rate budget for the current optimization window R_T is reached.

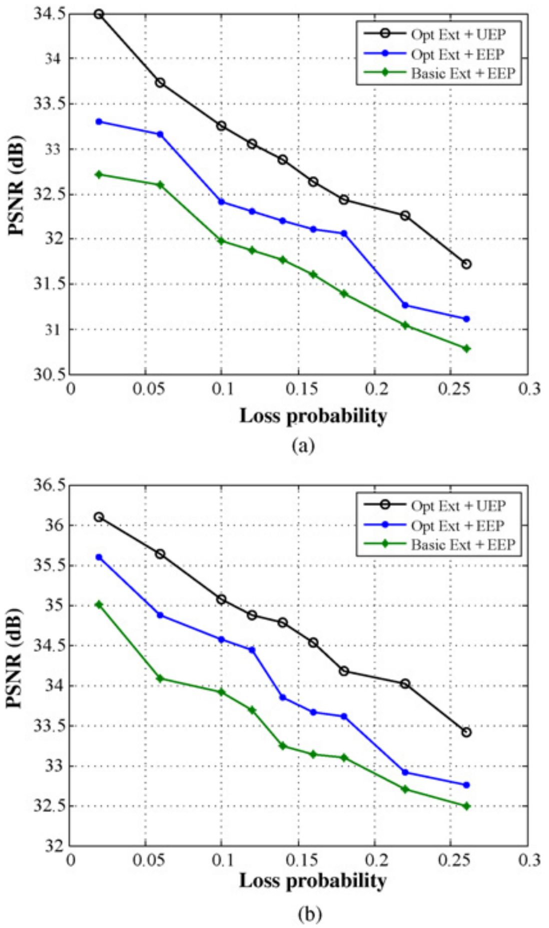


Fig. 7. PSNR performance of the three transmission systems versus packet loss rate. (a) *Stefan* QCIF, R_T = 500 kb/s. (b) *Coastguard* QCIF, R_T = 400 kb/s.

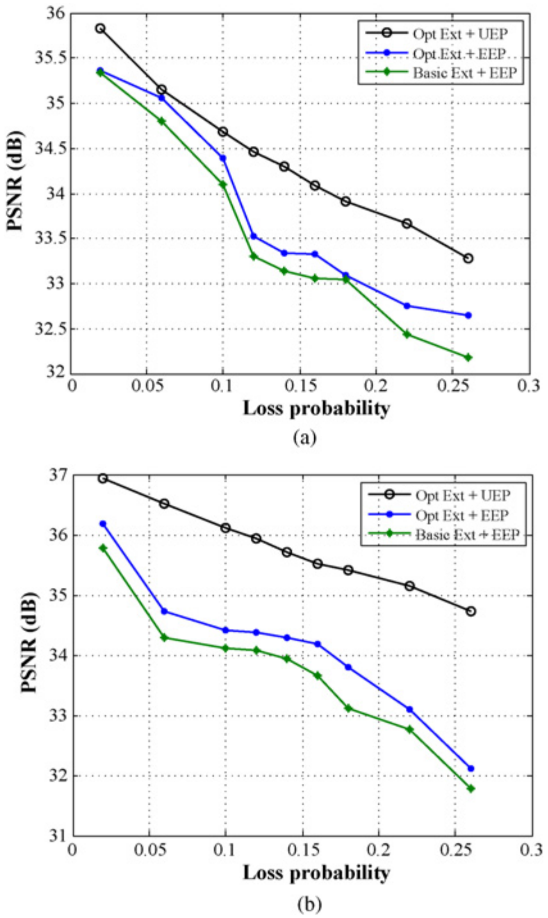
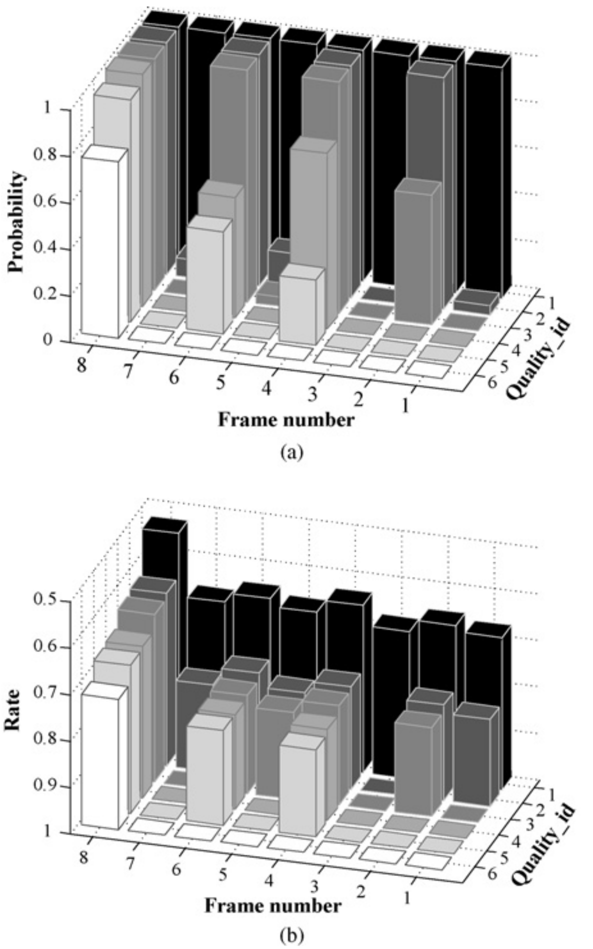


Fig. 8. PSNR performance of the three transmission systems versus packet loss rate. (a) *Tempete* CIF, R_T = 2 Mb/s. (b) *City* CIF, R_T = 900 kb/s.

V. Experimental Results

In this section, we evaluate the performance of our proposed optimized bit extraction and channel coding scheme for the H.264/AVC scalable extension. The simulation is implemented with the reference software joint scalable video model (JSVM) 9.15. Various test sequences at CIF and quarter common intermediate format (QCIF) resolutions are considered in our experiments. These sequences are encoded into two layers, a base layer and a quality layer, with basis quantization parameters $QP = 36$ and $QP = 24$, respectively. Furthermore, the quality layer is divided into five MGS layers. In our experiments, we used RS codes of the form $(32, k)$ with a symbol length of $m = 5$. All results were obtained using a 100 channel realization. We also assumed an i.i.d. channel model: each transport packet may be lost in the channel with a fixed loss probability, ϵ, independent of the others.

To evaluate the performance of the proposed UEP scheme, we consider a memoryless channel with various transport layer packet loss probabilities denoted by ϵ. Figs. 7 and 8 show the average peak signal-to-noise ratio (PSNR) of the decoded sequence for various test sequences/resolutions. The three transmission schemes considered here are: 1) our proposed join extraction with UEP, referred to as “Opt Extraction + UEP”; 2) our proposed source extraction with the best fixed channel coding rate obtained exhaustively from the set of channel coding rates for each transmission bit rate, referred to as “Opt Extraction + EEP”; and 3) JSVM basic extraction with the best fixed channel coding rate. In order to build fair comparison criteria, we assume that the base layers of the key frames are coded using the lowest channel coding rate and therefore always received intact for all three schemes. As illustrated in Figs. 7 and 8, the joint extraction with UEP outperforms the other two schemes. Note that packets in equal error protection schemes may be lost with a constant probability; however, the UEP scheme distributes parity bits such that important packets have smaller loss probabilities and therefore some less important packets have higher loss probabilities. Fig. 9 illustrates the allocation of the available bandwidth in the proposed system for the *City* CIF sequence. Fig. 9(a) shows the probability that a NAL unit in a particular location of a GOP is included in the queue for transmission. Fig. 9(b) on the other hand shows the average channel coding rates allocated to the NAL units when they are included for transmission. As expected, NAL units that belong to the higher levels in the prediction hierarchy are more often included in the transmission queue with adequate channel protection. Nevertheless, the actual rate allocation depends on the particular content of the GOP, as well as the available bandwidth and channel conditions.



(a)

(b)

Fig. 9. Bandwidth allocation of the proposed system, $\epsilon = 10\%$ and $R_T = 1$ Mb/s. (a) Probability of NAL unit inclusion. (b) Average channel coding rate allocated.

VI. Conclusion

A method was proposed for cross-layer optimization of the scalable extension of the H.264/AVC, which ensures robust delivery of scalable video over error-prone channels. The transmitter computes an estimate of the total distortion of the reconstructed frame at the decoder for the given available bandwidth, packet loss condition, and error concealment method. The algorithm recursively computes the total distortion of each GOP at a picture-level to accurately account for both enhancement layer clipping and drift propagation. The accuracy of the estimate was demonstrated via simulation results. We further incorporated the estimate within a cross-layer framework for optimized content-aware bit extraction and unequal channel protection. Using this framework, for a given transmission rate and channel condition, we identified packets with most expected contribution to the end video quality and their appropriate channel protection rate. Simulation results showed the effectiveness of the proposed framework compared to the JSVM content-independent bit extraction with equal error protection.

References

[1] *Amd.3 Scalable Video Coding*, document 14496-10.doc, Joint Draft ITU-T Rec. H.264/ISO/IEC, 2007.

[2] H. Schwarz, D. Marpe, and T. Wiegand, “Overview of the scalable video coding extension of the H.264/AVC standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[3] X. K. Ang, C. Zhu, Z. G. Li, X. Lin, G. N. Feng, S. Wu, and N. Ling, “Unequal loss protection for robust transmission of motion compensated video over the Internet,” *Signal Process. Image Commun.*, vol. 18, no. 2, pp. 157–167, Mar. 2003.

[4] F. Marx and J. Farah, “A novel approach to achieve unequal error protection for video transmission over 3G wireless networks,” *Signal Process. Image Commun.*, vol. 19, pp. 313–323, Apr. 2004.

[5] T. Fang and L. P. Chau, “A novel unequal error protection approach for error resilient video transmission,” in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 4. 2005, pp. 4022–4025.

[6] H. Cai, B. Zeng, G. Shen, and S. Li, “Error-resilient unequal protection of fine granularity scalable video bitstreams,” in *Proc. IEEE Int. Conf. Commun.*, vol. 3. Jun. 2004, pp. 1303–1307.

[7] M. van der Schaar and H. Radha, “Unequal packet loss resilience for fine-granular-scalability video,” *IEEE Trans. Multimedia*, vol. 3, no. 4, pp. 381–393, Dec. 2001.

[8] C. E. Costa, Y. Eisenberg, F. Zhai, and A. K. Katsaggelos, “Energy efficient wireless transmission of MPEG-4 fine granular scalable video,” in *Proc. IEEE Int. Conf. Commun.*, vol. 5. Jun. 2004, pp. 3096–3100.

[9] T. Fang and L. P. Chau, “GOP based channel rate allocation using genetic algorithm for scalable video streaming over error-prone networks,” *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1323–1330, Jun. 2006.

[10] D. E. Goldberg, “Computer implementation of a genetic algorithm,” in *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Jan. 1989, pp. 60–80.

[11] M. Stoufs, A. Munteanu, J. Cornelis, and P. Schelkens, “Scalable joint source-channel coding for the scalable extension of H.264/AVC,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 12, pp. 1657–1670, Dec. 2008.

[12] Z. Wu, A. Bilgin, and M. Marcellin, “Joint source/channel coding for image transmission with JPEG2000 over memoryless channels,” *IEEE Trans. Image Process.*, vol. 14, no. 8, pp. 1020–1032, Aug. 2005.

[13] Y. P. Fallah, H. Mansour, S. Khan, P. Nasiopoulos, and H. M. Alnuweiri, “A link adaptation scheme for efficient transmission of H.264 scalable video over multirate WLANs,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 7, pp. 875–887, Jul. 2008.

[14] M. Jubran, M. Bansal, L. 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. 18, no. 1, pp. 106–116, Jan. 2009.

[15] A. Majumda, D. Sachs, I. Kozintsev, K. Ramchandran, and M. Yeung, “Multicast and unicast real-time video streaming over wireless LANs,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 524–534, Jun. 2002.

[16] H. Schwarz, D. Marpe, and T. Wiegand, “Analysis of hierarchical B pictures and MCTF,” in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2006, pp. 1929–1932.

[17] *CE1: Simplified FGS*, document JVT-W090.doc, Joint Video Team, Apr. 2007.

[18] I. Amonou, N. Cammas, S. Kervadec, and S. Pateux, “Optimized rate-distortion extraction with quality layers in the scalable extension of H.264/AVC,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1186–1193, Sep. 2007.

[19] W.-H. Peng, J. K. Zao, H.-T. Huang, T.-W. Wang, and L.-S. Huang, “A rate-distortion optimization model for SVC inter-layer encoding and bitstream extraction,” *J. Vis. Commun. Image Represent.*, vol. 19, no. 8, pp. 543–557, 2008.

[20] E. Maani and A. K. Katsaggelos, “Optimized bit extraction using distortion modeling in the scalable extension of H.264/AVC,” *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 2022–2029, Sep. 2009.

Ehsan Maani (S’05) received the B.S. degree in physics from Sharif University of Technology, Tehran, Iran, in 2004, and the M.S. degree in electrical engineering from Northwestern University, Evanston, IL, in 2007. He is currently pursuing the Ph.D. degree in electrical engineering in the area of signal processing at the School of Electrical Engineering and Computer Science, Northwestern University.

His research interests include image/video compression and transmission, indexing, and retrieval.

Aggelos K. Katsaggelos (S'80-M'85-SM'92-F'98) received the Diploma in electrical and mechanical engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, in 1979, and the M.S. and Ph.D. degrees, both in electrical engineering, from the Georgia Institute of Technology, Atlanta, in 1981 and 1985, respectively.

In 1985, he was with the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, where he is currently a Professor. He was the Holder of the Ameritech Chair of Information Technology from 1997 to 2003. He is also the Director of the Motorola Center for Seamless Communications, a Member of the Academic Affiliate Staff, Department of Medicine, Evanston Hospital, Evanston, IL, and a Special Term Appointee at Argonne National Laboratory, Argonne, IL. He has published extensively in the areas of signal processing, multimedia transmission, and computer vision. He is the editor of *Digital Image Restoration* (Springer-Verlag, 1991), the coauthor of *Rate-Distortion Based Video Compression* (Kluwer, 1997), *Super-Resolution for Images and Video* (Claypool, 2007), *Joint Source-Channel Video Transmission* (Claypool, 2007), and the co-editor of *Recovery Techniques for Image and Video Compression and Transmission* (Kluwer, 1998). He is the co-inventor of 14 international patents.

Dr. Katsaggelos was the Editor-in-Chief of the IEEE SIGNAL PROCESSING MAGAZINE from 1997 to 2002, a Member of the Board of Governors of the IEEE Signal Processing Society from 1999 to 2001, and a Member of the Publication Board of the IEEE PROCEEDINGS from 2003 to 2007. He is a Fellow of the International Society for Optical Engineers. He was the recipient of the IEEE Third Millennium Medal in 2000, the IEEE Signal Processing Society Meritorious Service Award in 2001, the IEEE Signal Processing Society Best Paper Award in 2001, the IEEE International Conference on Multimedia and Expo Paper Award in 2006, and the IEEE International Conference on Image Processing Paper Award in 2007. He was a Distinguished Lecturer of the IEEE Signal Processing Society from 2007 to 2008.