

TWO-DIMENSIONAL CHANNEL CODING FOR SCALABLE H.264/AVC VIDEO

Ehsan Maani and Aggelos K. Katsaggelos

Northwestern University
Electrical Engineering and Computer Science
Evanston, IL

ABSTRACT

Efficient bit stream adaptation and resilience to packet losses are two critical requirements in scalable video coding for transmission over packet-lossy networks. These requirements have a greater significance in scalable H.264/AVC video bit streams since missing refinement information in a layer propagates to all lower layers in the prediction hierarchy and causes substantial degradation in video quality. This work proposes an algorithm to accurately estimate the overall distortion of the reconstructed frames due to enhancement layer truncation, drift/error propagation, and error concealment in the scalable H.264/AVC video. 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 PSNR compared to a conventional content independent system.

Index Terms— Channel coding, UEP, scalable video coding, H.264/AVC.

1. INTRODUCTION

Multimedia applications involving the transmission of video over communication networks are rapidly increasing in popularity. However, today's communication networks 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. The new SVC standard [1] which was approved as Amendment 3 of the Advanced Video Coding (AVC) standard, provides significantly higher compression efficiency compared to the scalable profile of the prior video coding standards. 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 dependency identifier D or temporal identifier T . Moreover, quality refinement layers inside each dependency layer are identified by a quality identifier Q . 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.

To transmit video efficiently and minimize the distortion of the reconstructed video sequence, the system resources should be properly allocated. Joint bit stream adaptation and unequal error protection (UEP) has been shown to be very effective in such environments since scalable video usually contains various parts with significantly different impact on the decoded signal quality. Bit stream adaptation refers to deliberately discarding a number of Network Adaptation Layer (NAL) units at the transmitter or in the network such that a particular average bit rate and/or resolution is reached. 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 dependencies. Many researchers have tackled this problem by applying UEP only to one dimension: either different frame types, e.g., [3] or to various scalable layers, e.g., [4]. Very few research works jointly consider these two aspects of the scalable video (a 2D UEP technique), e.g., [5]. However, these algorithms designed for MPEG-4 fine-granular-scalable (FGS) 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. 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 GOP except for the *key frames*. 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. 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.

2. PROBLEM FORMULATION

2.1. System Model

In this work, we consider a single-resolution SVC stream. Nonetheless, our calculations can be directly applied to the more general multi-resolution case if we assume that all quality NAL units associated with lower resolution spatial layers are included before the base

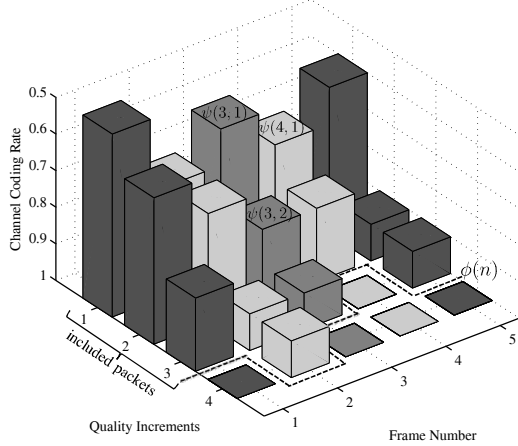


Fig. 1. Example of a selection map and channel rate allocation for a single resolution bit stream.

quality of a higher resolution. Let $\pi(n, q)$ represent the NAL unit associated with frame n and quality level q ($q = 0$ represents the base quality). Since each NAL unit $\pi(n, q)$ is useful at the decoder only if $\pi(n, q - 1)$ is available, we can denote a subset of the NAL units by a *selection map* $\phi : \mathbb{Z}^+ \rightarrow \mathbb{Z}^+$ such that $\phi(n)$ represents the number of quality NAL units included in the selection for each frame n . Note that $\phi(n) = 0$ indicates that no NAL unit (including the base NAL) for frame n has been included in the set. Therefore, frames which depend on it through motion compensated prediction (MCP) are also undecodable. We further define the channel coding function for each NAL unit $\psi : \mathbb{Z}^{+2} \rightarrow (0, 1]$ such that $\psi(n, q)$ denotes the channel rate allocation associated with $\pi(n, q)$. Then, the problem of optimal bit extraction and channel rate allocation can be formulated as

$$\begin{aligned} (\phi^*, \psi^*) &= \min_{\phi \in \Phi, \psi \in \Psi} E\{D(\phi, \psi; \epsilon)\} \\ \text{s.t. } R(\phi, \psi) &\leq R_T, \end{aligned} \quad (1)$$

where ϕ and ψ are vector representations of the ϕ and ψ functions respectively with element values of $\phi(n)$ and $\psi(n, q)$ for all n and q , respectively. Ψ is the set of all possible channel coding rates, R_T is the total transmission rate and ϵ represents the transport layer packet loss probability. 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 $\phi(n)$ and the associated channel coding rates $\psi(n, q)$ as well as the transport packet loss probability. An example of selection and channel coding rate functions is illustrated in Fig. 1.

Due to the huge complexity of the optimization in (1), a solution cannot be found using a simple non-linear optimization scheme. For a given packet selection, 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. Therefore, a fast yet accurate approximation of the expected distortion is critical in solving the optimization in (1).

3. EXPECTED DISTORTION ESTIMATION

In this section we develop an approximation method for the computation of the expected distortion. Our expected distortion model assumes knowledge of Channel State Information (CSI) and the particular error concealment method employed by the decoder. In this work a simple and popular concealment strategy is employed: the lost picture is replaced by the nearest temporal neighboring picture. We consider a generic case where a packet loss probability of p_n^q is assigned to the q^{th} quality increment packet of frame n , i.e., $\pi(n, q)$. Recall that p_n^q is dependant on the transport packet loss probability and the specific channel coding rate $\psi(n, q)$. Additionally, let the set $\mathcal{S} = \{s_0, s_1, \dots, s_N\}$ represent the N pictures in the GOP plus the key picture of the preceding GOP denoted by s_0 . Note that in our notation the n^{th} 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

$$\begin{aligned} E\{\tilde{D}_n|BL\} &= \sum_{q=1}^{\phi(n)} p_n^q D_n(q-1) \prod_{i=0}^{q-1} (1-p_n^i) + \\ &D_n(\phi(n)) \prod_{i=0}^{\phi(n)} (1-p_n^i), \end{aligned} \quad (2)$$

where $D_n(q)$ is the total distortion of frame n reconstructed by inclusion of $q > 0$ quality increments. The first term in equation (2) accounts for cases in which, all $(q-1)$ quality segments have been successfully received but the q^{th} 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.

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. 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 $D_{n,k}^{\text{con}}$. Consequently, the expected distortion of frame n is computed according to

$$\begin{aligned} E\{\tilde{D}_n\} &= \sum_{i \in \Delta_n} p_i^0 D_{n,k}^{\text{con}} \prod_{\substack{j \in \Delta_n \\ j \prec i}} (1-p_j^0) + \\ E\{\tilde{D}_n|BL\} &\prod_{j \in \Delta_n} (1-p_j^0), \end{aligned} \quad (3)$$

where k represents the concealing frame. The first term in equation (3) 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 con-

cealed using a decodable temporal neighbor while the second term indicates the case in which all base layers are received and thus the frame quality we expect is $E\{\tilde{D}_n|BL\}$.

From equations (3) and (2), 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 n^{th} 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. Generally, two different sources of distortion contribute to $D_n(q)$: distortion propagated from the ancestor frames and the distortion caused by missing quality increments of the current frame n . We refer to these distortions as drift and EL truncation distortions denoted by D_n^d and $D_n^e(q)$, respectively. In [6] we explained that if both D_n^d and $D_n^e(q)$ are known, the total distortion can be estimated as

$$D_n(q) \approx D_n^d + D_n^e(q) + 2\kappa\sqrt{D_n^d}\sqrt{D_n^e(q)}, \quad (4)$$

where κ is a constant in the range $0 \leq \kappa \leq 1$ obtained experimentally from test sequences (a typical value $\kappa = 0.05$). Fortunately, the error due to EL truncation, $D_n^e(q)$, 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. As shown in [6], the drift distortion can be estimated from the total distortion of the parent frames. Nevertheless, the exact distortion of the parent frame is unknown due to indeterministic nature of the packet losses in the channel. Consequentially, the expected distortion value of the parent frames is employed to estimate the drift caused in the child frame, i.e.,

$$D_n^d \approx \sum_{i \in \Lambda_n} \alpha_i E\{\tilde{D}_i|BL\} + \sum_{i \in \Lambda_n} \sum_{j \in \Lambda_n} \beta_{ij} E\{\tilde{D}_i|BL\} E\{\tilde{D}_j|BL\}, \quad (5)$$

where Λ_n denotes the set of the two parents of frame n . α_i and $\beta_{i,j}$ are constant numbers for each frame derived by evaluating the sequence quality when some of the NAL units are excluded [6]. Similarly, when the base layer is missing or undecodable, the distortion after concealment is estimated based on the quality of the frame used for the concealment as

$$D_{n,k}^{con} \approx \mu_k + \nu_k E\{\tilde{D}_k|BL\}, \quad (6)$$

where μ_k and ν_k are constants obtained as described in [6] for each concealing option k .

The total distortion $D_n(q)$ is then computed according to equation (4). 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 equation (2). This value is then used to find the drift distortion of the child frame utilizing equation (5). This drift distortion then yields to the com-

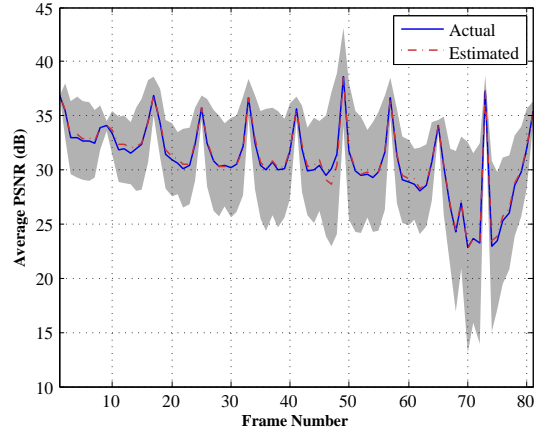


Fig. 2. Actual versus estimated frame distortions for foreman CIF sequence; packet loss rate $p = 10\%$.

putations of $D_n(q)$ and $E\{\tilde{D}_n|BL\}$ for the child frame according to equations (4) and (2), 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. From equation (3), it is apparent that one can obtain the total expected distortion for the entire GOP once the values of $E\{\tilde{D}_n|BL\}$ and $D_{n,k}^{con}$ are known for all n . To evaluate the accuracy of the proposed model, we compared the calculated expected distortion to an average of the decoded distortion for various loss patterns. Figure 2 shows an example of this comparison for the Foreman 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 probability (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. Notice that the grey area indicates the standard deviation of the reconstructed signal quality over all channel realizations.

4. SOLUTION ALGORITHM

The distortion model proposed in this work 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 the above mentioned 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 equations (2) and (3), the expected distortion of the video sequence directly depends on the source mapping function $\phi(n)$. Its dependency 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.

The optimization can be performed over an arbitrary number of GOPs, denoted by M . The source mapping function $\phi(n)$ initially only includes the base layer of the key pictures with an initial channel coding rate less than 1. Then, at each time step, a decision is made whether to add a new packet to the transmission queue or increase the FEC 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

$$\delta ED^* = \max_n \max_{q < \phi(n)} \left| \frac{\partial ED(\phi, \psi) / \partial \psi(n, q)}{\partial R_t(\phi, \psi) / \partial \psi(n, q)} \right|, \quad (7)$$

where ED and R_t represent the expected distortion and the total rate associated with the current ϕ and ψ . 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, δED^\dagger , i.e.,

$$\delta ED^\dagger = \max_n \max_{\psi(n, q) \in \Psi} \left| \frac{\partial^2 ED(\phi, \psi) / \partial \phi(n) \partial \psi(n, q)}{\partial^2 R_t(\phi, \psi) / \partial \phi(n) \partial \psi(n, q)} \right|, \quad (8)$$

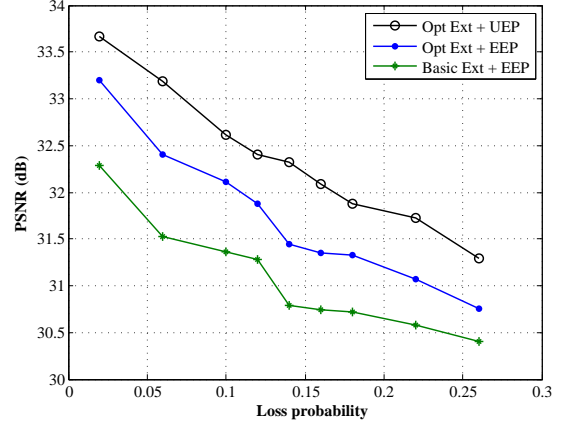
where $q = \phi(n)$. In cases for which $\delta ED^* > \delta ED^\dagger$, the channel protection rate of the already included packet $\pi(n^*, q^*)$ is incremented to the next level by padding additional parity bits. Conversely, when $\delta ED^* < \delta ED^\dagger$, the source packet $\pi(n^\dagger, \phi(n^\dagger))$ is included in the transmission queue with a channel coding rate $\psi(n^\dagger, \phi(n^\dagger))$ obtained from equation (8). Note that in both scenarios, the corresponding functions ϕ and ψ 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.

5. EXPERIMENTAL RESULTS

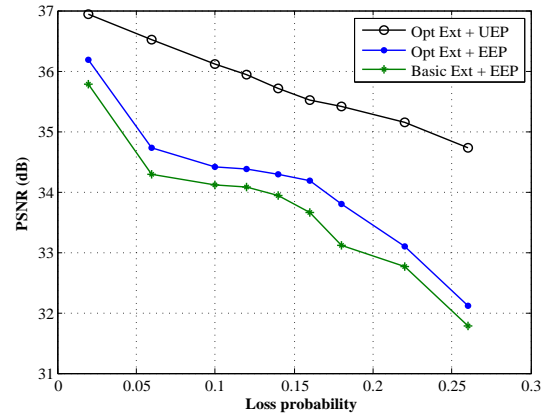
In this section, we evaluate the performance of our proposed optimized bit extraction and channel coding scheme for the H.264/AVC scalable extension. The simulation is implemented with the reference software JSVM 9.15. These sequences are encoded into two layers, a base layer and a quality layer, with basis quantization parameters $QP = 36$ and $QP = 24$, respectively. Furthermore, the quality layer is divided into 5 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 100 channel realizations. To evaluate the performance of the proposed UEP scheme, we consider a memoryless channel with various transport layer packet loss probabilities denoted by ϵ . Figure 3 shows the average PSNR of the decoded sequence for two different sequences/resolutions. The three transmission schemes considered here are: 1) Our proposed joint 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’’; 3) JSVM basic extraction with the best fixed channel coding rate. For demonstration purposes, 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 Fig. 3 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.

6. REFERENCES

[1] ‘‘Joint draft ITU-T rec. H.264 — ISO/IEC 14496-10 / amd.3 scalable video coding,’’ 2007.



(a) Mobile QCIF



(b) City CIF

Fig. 3. PSNR performance of the three transmission systems versus packet loss rate (a) Mobile QCIF, $R_T = 500$ kbps (b) City CIF, $R_T = 900$ kbps.

[2] H. Schwarz, D. Marpe, and T. Wiegand, ‘‘Overview of the scalable video coding extension of the h.264/avc standard,’’ *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 9, pp. 1103–1120, Sept. 2007.

[3] H. L. Huang and S. Liang, ‘‘Unequal error protection for MPEG-2 video transmission over wireless channels,’’ *Signal Processing: Image Communication*, vol. 19, no. 1, pp. 67–79, Jan. 2004.

[4] M. van der Schaar and H. Radha, ‘‘Unequal packet loss resilience for fine-granular-scalability video,’’ vol. 3, no. 4, pp. 381–393, Dec. 2001.

[5] T. Fang and L. P. Chau, ‘‘GOP based channel rate allocation using genetic algorithm for scalable video streaming over error-prone networks,’’ *Image Processing, IEEE Transactions on*, vol. 15, no. 6, pp. 1323–1330, June 2006.

[6] Ehsan Maani and Aggelos K. Katsaggelos, ‘‘Optimized bit extraction using distortion estimation in the scalable extension of h.264/avc,’’ in *IEEE International Symposium on Multimedia (ISM)*, Berkeley, CA, 2008.