Optimal Streaming of Layered Video: Joint Scheduling and Error Concealment

Philippe de Cuetos
Institut EURECOM
2229, route des Crêtes
06904 Sophia Antipolis, France
philippe.de-cuetos@eurecom.fr

Keith W. Ross
Polytechnic University
Six MetroTech Center
Brooklyn, NY 11201, USA
ross@poly.edu

ABSTRACT

We consider streaming layered video (live and stored) over a lossy packet network in order to maximize the video quality that is rendered at the receiver. We propose an end–to–end framework in which packet scheduling decisions at the sender explicitly account for the error concealment mechanism at the receiver. We refer to this framework as joint scheduling and error concealment. We show how the theory of infinite–horizon, average–reward Markov decision processes with average–cost constraints can be applied to the joint scheduling and error concealment problem. The formulation allows for a wide variety of performance metrics, including metrics that take quality variation into account. We demonstrate the framework and solution procedure using MPEG–4 FGS video traces.

1. INTRODUCTION

In this paper we consider streaming layered video over a lossy packet network in order to maximize the video quality that is rendered at the receiver. We propose an end–to–end framework in which packet scheduling decisions at the sender explicitly account for the error concealment mechanism at the receiver. We refer to this framework as *joint scheduling and error concealment*.

In many packet network environments, including the Internet, the bandwidth available to a streaming application is not known *a priori* and varies throughout the streaming application. For such network environments, layered–encoded video is appropriate [10, 12, 17, 18, 22]. The video is encoded into a Base Layer (BL) and a number of enhancement layers (ELs). The decoded BL provides minimal rendered quality; additional decoded ELs progressively enhance the rendered quality.

In a typical streaming application, the sender dynamically *schedules* the transmission of media packets as a function of available bandwidth in order to maximize the rendered video quality. The sender may choose not to transmit some media packets, thereby not sending some layers in some frames. In general, the scheduling may also include the retransmission of lost packets that can arrive at the receiver before their decoding deadlines. At the receiver, some of the media packets are available on time, that is, before their decoding deadlines. Other packets are not be available, either because they were transmitted and lost, or simply because the sender never scheduled them for transmission. At the time of rendering to the user, the decoder typically applies several methods of *error concealment* in order to best conceal the missing packets. Error Concealment (EC) consists in exploiting the spatial and temporal correlations of audio or video to interpolate missing packets from the surrounding available packets [21]. For video, a simple and popular method for temporal error concealment is to display, instead of the missing macro block from the current frame, the macro block at the same spatial location but from the previous frame.

Packet scheduling and error concealment are two fundamental components in an end–to–end video streaming system. Figure 1 illustrates their respective functions. At the sender, the scheduler determines the layers that should be sent to the receiver. At the receiver, before rendering the media, the decoder performs error concealment from the available layers. Traditionally, packet scheduling and error concealment are designed independently without considering any interplay between the two. In particular, the scheduling policy is normally optimized without taking into account the presence of error concealment at the receiver [6, 15, 16].

In this paper, we argue that the scheduling and error concealment components of a video streaming system should be designed jointly and not separately. In particular, when designing a scheduling policy, not only should we account for the layered structure of the media, the channel characteristics, and the effects of missing packets on distortion, but we should also explicitly account for error concealment at the receiver. Thus, we argue for a more unified, end–to–end approach for designing video streaming systems.

This paper has two main contributions. First, we present a new optimization framework for joint packet scheduling and temporal error concealment. Using MPEG–4 FGS video traces, we compare joint scheduling and error concealment to "disjoint" scheduling and error concealment:

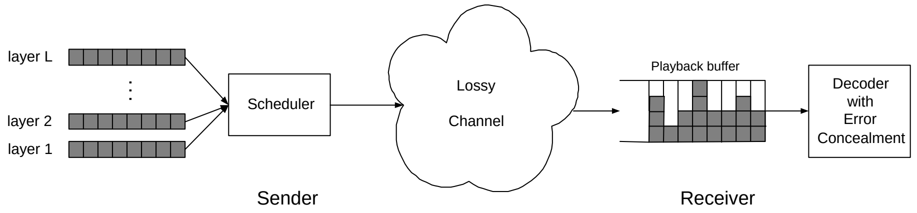


Figure 1: Video streaming system

- **Disjoint Scheduling and Error Concealment (Disjoint S+EC):** The sender determines and employs the optimal scheduling policy, which is obtained without accounting for error concealment at the receiver. The receiver nevertheless applies error concealment before rendering the video.
- **Joint Scheduling and Error Concealment (Joint S+EC):** The sender determines and employs the optimal scheduling policy, which accounts for error concealment. The receiver applies error concealment before rendering the video.

It is important to note that both schemes employ EC, so that when comparing the rendered video quality of the two schemes, we are indeed making a fair comparison. We show that for the same bandwidth, Joint S+EC can significantly improve video quality over Disjoint S+EC. (Conversely, it can be shown that with Joint S+EC, the streaming application can achieve the same overall video quality as Disjoint S+EC but with significantly less bandwidth.)

Our second main contribution is to show how the theory of infinite–horizon, average–reward Markov Decision Processes (MDPs) *with average–cost constraints* can be applied to the joint scheduling and error concealment problem. To our knowledge, infinite horizon constrained MDPs have not been applied yet to video streaming. We show how constrained MDPs can be used for a wide variety of quality metrics, including metrics which take quality variation into account. We also show that for streaming applications with small playout delays (such as live streaming), the constrained MDP approach is computationally tractable, providing optimal scheduling policies for the Joint S+EC scheme.

This paper is organized as follows. We conclude this section with a discussion of related work. In Section 2, we illustrate the potential benefits of Joint S+EC. In Section 3, we introduce a novel optimization framework for Joint S+EC. In Section 4, we show how constrained MDPs can be applied to the optimal Joint S+EC problem. We provide several numerical examples which show that Joint S+EC can provide significant improvements in performance over Disjoint S+EC. In Section 5, we consider a modified version of the problem, in which there is an additional constraint on quality variability. In Section 6, we provide simulation results for MPEG–4 FGS video traces, which confirm the benefits of Joint S+EC over Disjoint S+EC. We conclude in Section 7.

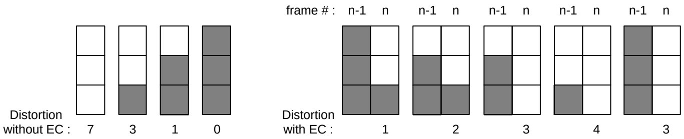


Figure 2: Example of expected distortion values for a video encoded with 3 layers.

1.1 Related Work

To our knowledge, the most closely related work to this paper is the work of Chou and Miao [5, 6] which considers rate–distortion optimized streaming. Chou and Miao consider scheduling packetized media over a packet erasure channel in order to minimize an additive combination of distortion and average rate. The series of papers makes a number of contributions, both in developing a novel optimization framework, and applying the framework to a variety of packet network models. The most important difference between their work and ours is that Chou and Miao do not consider error concealment in their optimization framework. We introduce the joint scheduling and error concealment problem. Also, Chou and Miao develop a heuristic algorithm for finding a sub–optimal scheduling policy, whose performance may be significantly below the truly optimal scheduling policy. Our constrained MDP approach provides a tractable means for determining the truly optimal Joint S+EC policy. (However, the framework of Chou and Miao allows for retransmissions, whereas we suppose no retransmissions.) Finally, the framework provided in this paper can handle quality variability metrics, and can be extended to handle error correction with FEC.

Other works on optimal scheduling of media using a feedback channel include [16, 20]. These works do not consider error concealment. Podolsky et al. [16] study optimal retransmission strategies of scalable media. Their analysis is based on Markov chains with a state space that grows exponentially with the number of layers. Servetto [20] studies scheduling of complete GOPs encoded in multiple description codes. The sender adapts the number of descriptions sent to the receiver, as a function of the network state which is modeled as a HMM.

2. BENEFITS OF JOINT S+EC

In this section we provide a simple example to highlight the benefits of joint scheduling and error concealment. Consider a video segment composed of five frames, each of which is encoded into three layers. We suppose in this example that each frame is independently encoded. The only dependen-

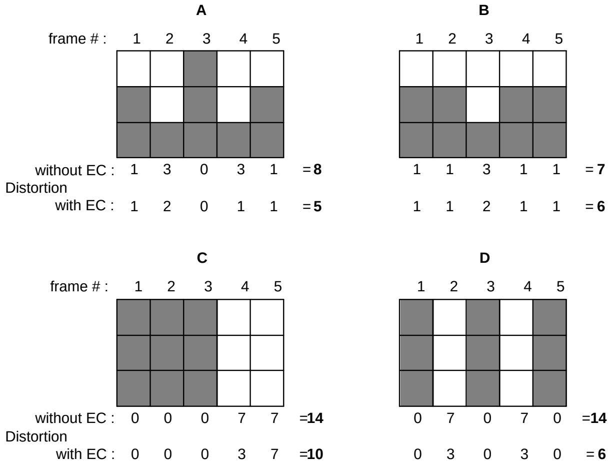


Figure 3: Example of scheduling policies transmitting 9 packets.

cies are due to layered encoding, i.e., a given layer of a video frame needs all the lower layers of the same frame to be decoded. We suppose that each layer fits exactly into one packet, all packets are the same size, and all frames have the same rate–distortion functions.

On the left of Figure 2, we give the distortion values for each frame, as a function of the number of layers which are available for the frame. These are distortion values expected by the sender without considering temporal EC at the receiver. On the right of Figure 2, we show the distortion values for frame n when EC from the previous frame $n-1$ is used. These are the distortion values which are actually obtained after decoding.

Figure 3 shows four possible scheduling policies at the sender (A, B, C, and D), when we require that each policy send exactly nine packets. Initially, suppose that there is no packet loss. For each scheduling policy, we give the total distortion with and without EC. Now consider the optimal Disjoint S+EC policy and the optimal Joint S+EC policy. The optimal Disjoint S+EC policy is policy B, which minimizes the distortion at the receiver without taking EC into account. After applying error concealment to policy B, the resulting distortion is 6. Hence, the optimal Disjoint S+EC policy has a distortion of 6. But the optimal Joint S+EC policy is policy A, which has a lower rendered distortion than policy B.

3. PROBLEM FORMULATION

In this paper, we consider video streaming, live or stored. The video at the sender is encoded into L layers of constant size. Recall that the main property of layered–encoded video is that layer l of a given frame can not be decoded unless all lower layers $1, \ldots, l-1$ are also available at the decoder.

We suppose that the video contains N frames, and that its L layers are not motion–compensated, i.e., the decoding of layer l of frame n does not depend on the decoding of previous frames $n-1, \ldots$ (this is the case for the FGS–EL defined in the MPEG–4 standard [1]). Also, we suppose that the additional quality brought by a given layer is roughly constant for all frames of the video (i.e., layer l of frame n brings roughly the same amount of quality to frame n than layer l of frame $n+1$ to frame $n+1$). More generally, for long videos containing multiple scenes with different visual characteristics, the quality brought by a layer is likely to vary significantly for different parts of the video [9]. In this case, we suppose that the video has been previously segmented into homogeneous segments of video frames, such that the quality brought by each layer is roughly constant throughout the segment. Therefore, in this study, we consider a single homogeneous segment containing N frames. In the case of longer videos, we would apply our optimization framework to each separate segment.

Throughout this paper, we suppose that the transmission channel is a packet–erasure channel with instant feedback. The channel has a probability of success of q. We do not consider channel error correction, such as Forward Error Correction (FEC) or selective retransmissions. However, our framework can accommodate FEC codes as additional layers (this will be the subject of future work).

At the decoder, we suppose that, in order to conceal loss of packets for frame number n, only information from previous frame $n-1$ is used. However, information from frame $n-1$ does not necessarily fully conceal loss of packets from frame n. Note that, in practice, information from a set of consecutive previous frames, and even from subsequent frames, can also be used to perform error concealment for the current frame at the decoder. This has the potential to increase the accuracy in predicting any missing packet, but at the cost of an increase in run–time complexity of the decoder [21]. The theory presented here can be extended to handle these more sophisticated forms of error concealment; in order to see the forest through the trees, throughout we focus on only using the previous frame in error concealment.

For a given scheduling policy σ, let $rate_{avg}(\sigma)$ denote the average transmission rate for the video, which is in units of number of transmitted layers. Let $dist_{avg}(\sigma)$ denote the average distortion of the rendered video after error concealment. A typical problem formulation of rate–distortion optimized streaming is the following [6, 22]:

Problem 1. *Find an optimal scheduling policy σ which minimizes $dist_{avg}(\sigma)$ subject to $rate_{avg}(\sigma) \leq \alpha$,*

where α is the maximum transmission rate which is allowed by the network connection, or alternatively, the rate budget which is allocated to the streaming.

It may be misleading to solely use average image distortion, usually expressed in terms of average MSE (Mean Squared Error), to account for the quality of the rendered video. First, the average image distortion does not measure temporal artifacts, such as mosquito noise (moving artifacts around edges) or drifts (moving propagation of prediction errors after transmission). Second, high variations in quality between successive images may decrease the overall perceptual quality of the video. Therefore, the formulation of our problem should incorporate additional quality constraints. In this paper, we treat as an example the case of variations in quality between consecutive images. For a given scheduling policy σ, let $var_{avg}(\sigma)$ denote the average variation in distortion between two consecutive images. We can now formulate the following problem:

PROBLEM 2. *Find an optimal scheduling policy σ which minimizes $dist_{avg}(\sigma)$ subject to $rate_{avg}(\sigma) \leq \alpha$ and $var_{avg}(\sigma) \leq \gamma$,*

where γ is the maximum average variation in distortion which is allowed. (Its value can be found from subjective tests.)

Let $\mathcal{L} := \{0, 1, \ldots, L\}$. Let $A_n \in \mathcal{L}$ denote the scheduling action that the sender takes for frame n, i.e., the number of successive layers to send to the receiver for frame n. Note that transmitting layer l of a given frame without transmitting layer $l-1$ of that frame makes little sense. Indeed, layer l will never be decoded if the sender has not sent layer $l-1$. Let $X_n \in \mathcal{L}$ denote the state at the receiver for previous frame $n-1$, i.e., the number of successive layers which are available at the decoder for frame $n-1$. The order of succeeding actions and states is $X_{n-1}, A_{n-1}, X_n, A_n, \ldots$. Let D_n denote the distortion of frame n after decoding.

Throughout this paper we suppose that the sender can observe X_n when choosing the action A_n. This implies a feedback channel from receiver to sender with a RTT that is less than one frame time. The model can also be extended to handle delayed feedback, which will be the subject of a future paper. Also, our system does not allow for retransmission of lost packets. This is a reasonable assumption for live streaming. It is also reasonable for stored video systems with short playback delays and high VCR–like interactivity.

We denote by d_l, the distortion of a frame containing only the first l layers before temporal EC. (Without loss of generality, we take $d_L = 0$ and $d_0 = 1$.) We have $d_L < d_{L-1} < \cdots < d_1 < d_0$. For $0 \leq i, j \leq L$, we denote by d_{ij} the distortion of a frame after temporal error concealment, when i layers of the previous frame and j layers of the current frame were received by the decoder. Whenever $i \leq j$, the decoder cannot conceal lost layers of the current frame from the previous frame, therefore $d_{ij} = d_j$. We denote by *distortion matrix*, matrix $[d_{ij}]_{0 \leq i,j \leq L}$.

In our system, we suppose that the sender knows the distortion matrix of the current video segment. When streaming stored video, the distortion matrix can be computed off–line from the original uncompressed video segment. It can be stored at the sender, together with the video file. When streaming live video, the sender needs to estimate the value of the distortion matrix before starting the encoding and transmission of the current video segment. This estimate can be based on the previous video segments which have been encoded and already sent to the receivers. Since in most applications of live video streaming, such as streaming of sport events or videoconferencing, the consecutive video segments have usually recurrent or similar characteristics, we expect that the estimation of the distortion matrix of an upcoming segment can be made sufficiently accurate.

4. JOINT S+EC OPTIMIZATION

In this section we study Problem 1. We show that Problem 1 can be formulated as a constrained MDP, which can in turn be solved by linear programming [11, 13]. The problem is naturally formulated as a finite–horizon MDP with N steps, where N is the number of frames in a video segment. However, the computational effort associated with a finite–horizon MDP can be costly when N is large [3]. This may be a serious impediment for real–time senders. Therefore, we instead use infinite–horizon constrained MDPs. They have optimal stationary policies and have lower computational cost. The infinite horizon assumption corresponds to considering infinite–length video segments ($N = \infty$). Throughout this study, the values $rate_{avg}(\sigma)$, $distort_{avg}(\sigma)$ and $var_{avg}(\sigma)$ will be long–run averages.

4.1 Analysis

We consider the Markov Decision Process $\{X_n, A_n, n = 0, \ldots\}$. We define the reward associated with action A_n, when the receiver state is X_n, as:

$$r_n(X_n, A_n) = -E[D_n | X_n, A_n], \tag{1}$$

and the cost of choosing action A_n at transmission slot n as:

$$c_n(X_n, A_n) = A_n. \tag{2}$$

From these definitions, and given that $E[r_n(X_n, A_n)] = -E[D_n]$, Problem 1 can be rewritten as finding an optimal policy σ^* which maximizes the long–run average reward:

$$\lim_{n\to\infty} \frac{1}{n} E_\sigma[\sum_{m=1}^{n} r_m(X_m, A_m)] \text{ s.t. } \lim_{n\to\infty} \frac{1}{n} E_\sigma[\sum_{m=1}^{n} c_m(X_m, A_m)] \leq \alpha, \tag{3}$$

which falls into the general theory of constrained MDPs.

We can calculate the reward and cost as follows:

$$r_n(i, a) = -(1-q)(\sum_{j=0}^{a-1} d_{ij} q^j) - d_{ia} q^a \tag{4}$$

$$c_n(i, a) = a \tag{5}$$

For a randomized stationary policy σ, let $\sigma_{ia} = P_\sigma(A_n = a | X_n = i)$. We denote by $P_{iaj} = P(X_{m+1} = j | X_m = i, A_m = a)$ for the law of motion of the MDP. It is given by:

$$P_{iaj} = \begin{cases} 0 & \text{whenever } j > a \\ q^a & \text{when } j = a \\ q^j(1-q) & \text{otherwise.} \end{cases} \tag{6}$$

This MDP is clearly a unichain MDP [11, 19]. It therefore follows that the optimal policy for the constrained MDP is a randomized stationary policy. Furthermore, randomization occurs in at most one state [19]. An optimal stationary policy σ^* may be obtained from the following procedure:

Step 1. Find an optimal solution $z^* = \{z^*_{ia}, (i, a) \in \mathcal{L}^2\}$ to the linear program (LP):

$$\max \sum_{i\in\mathcal{L}}\sum_{a\in\mathcal{L}} r(i,a)z_{ia} \text{ s.t.} \begin{cases} \sum_{i\in\mathcal{L}}\sum_{a\in\mathcal{L}} c(i,a)z_{ia} \leq \alpha \\ \sum_{i\in\mathcal{L}}\sum_{a\in\mathcal{L}}(\delta_{ij} - P_{iaj})z_{ia} = 0, j \in \mathcal{L} \\ \sum_{i\in\mathcal{L}}\sum_{a\in\mathcal{L}} z_{ia} = 1 \\ z_{ia} \geq 0, i \in \mathcal{L}, a \in \mathcal{L}. \end{cases} \tag{7}$$

Let $\mathcal{L}^* = \{i \in \mathcal{L} : z^*_{ia} > 0 \text{ for some } a \in \mathcal{L}\}$.

Step 2. Determine an optimal policy σ^* as follows:

$$\begin{cases} \text{for } i \in \mathcal{L}^*, & \sigma^*_{ia} = \frac{z^*_{ia}}{\sum_{a\in\mathcal{L}} z^*_{ia}} \\ \text{for } i \notin \mathcal{L}^*, & \sigma^*_{ia} = 1 \text{ for some arbitrary } a \in \mathcal{L}. \end{cases} \tag{8}$$

Note that there are several algorithms to solve LPs. The most popular is the simplex algorithm. It has exponential worst–case complexity, but requires a small number of iterations in practice. There are other more elaborate algorithms which have polynomial complexity, such as the projective algorithm by Karmarkar [14]. LP (7) for the infinite horizon problem consists of $(L+1)^2$ variables and at most $L+3$ constraints. The corresponding LP formulation for the constrained MDP with finite horizon N ($N < \infty$) would consist into $N * (L+1)^2$ variables and $N * (L+1) + 1$ constraints, which is likely to increase the computational time significantly for high values of N.

4.2 1 Layer Video

As an example, consider the particular case with 1 layer ($L = 1$): scheduling consists in deciding at each decision epoch whether to send the single layer or send nothing at all. For this special case, we can actually derive a closed–form expression for the optimal policy (thereby circumventing linear programming). After analysis, the Joint S+EC optimal policy can be expressed as:

$$\begin{cases} \sigma^*_{01} = \alpha/(1-\alpha q),\ \sigma^*_{11} = 0 & \text{if } \alpha \leq 1/(1+q), \\ \sigma^*_{01} = 1,\ \sigma^*_{11} = 1 + (\alpha - 1)/(\alpha q) & \text{otherwise.} \end{cases} \tag{9}$$

The optimal average transmission rate and distortion are given by:

$$rate^*_{avg} = \alpha \tag{10}$$

$$dist^*_{avg} = \begin{cases} 1 - \alpha q(2 - d_{10}) & \text{if } \alpha \leq 1/(1+q), \\ (1-\alpha q)(1 - q(1-d_{10})) & \text{otherwise.} \end{cases} \tag{11}$$

In order to illustrate these results, consider the following distortion matrix:

$$[d_{ij}] = \begin{bmatrix} 1 & 0 \\ 0.5 & 0 \end{bmatrix} \tag{12}$$

In Figure 4, we show the minimum average distortion $dist^*_{avg}$ as a function of the maximum average transmission rate α, for selected values of the channel success rate q. We observe that the difference between the values of $dist^*_{avg}$ for different channel success rates increases with α. Indeed, for low values of α, optimal policies are likely to send very few frames ($\sigma^*_{01} \approx 0$ and $\sigma^*_{11} = 0$), so channel losses do not have much effect. However, for higher values of α, optimal policies send a large number of frames ($\sigma^*_{01} = 1$ and $\sigma^*_{11} \approx 1$), so the value of the channel success rate q has higher importance.

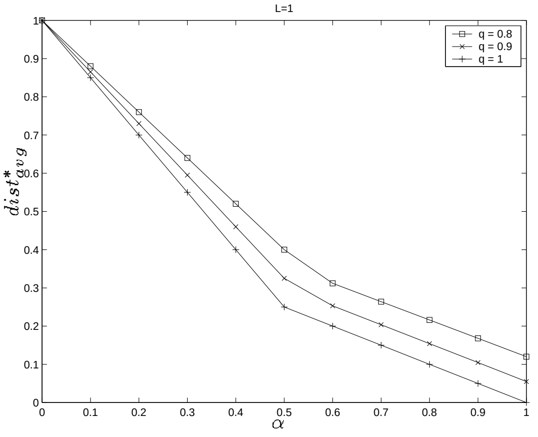


Figure 4: Minimum average distortion for $L = 1$.

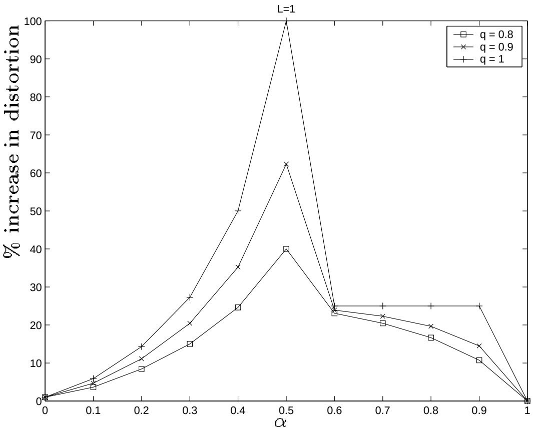


Figure 5: Worst–case increase in distortion of Disjoint S+EC policy, for $L = 1$.

To obtain the optimal policy for the case of Disjoint S+EC, we solve LP (7) with $d_{10} = 1$. We find that there is an infinite number of optimal transmission policies, including the previous optimal policy for Joint S+EC. The minimum expected distortion without considering EC is given by $dist^*_{avg} = 1 - \alpha q$; however, the achieved distortion after EC depends on the chosen optimal policy. In Figure 5, we give the worst–case increase in distortion after using a Disjoint S+EC optimal policy instead of the Joint S+EC optimal policy. We see that using Disjoint S+EC optimal policies can be highly detrimental to the rendered quality of the video, especially for medium values of α and for high channel success rates (this has been verified for values of d_{10} other than 0.5).

4.3 General Case of Multiple Layers

In the general case of multiple layers, we use LP (7) to solve Prob 1. We consider the example of 3 layers given in Fig-

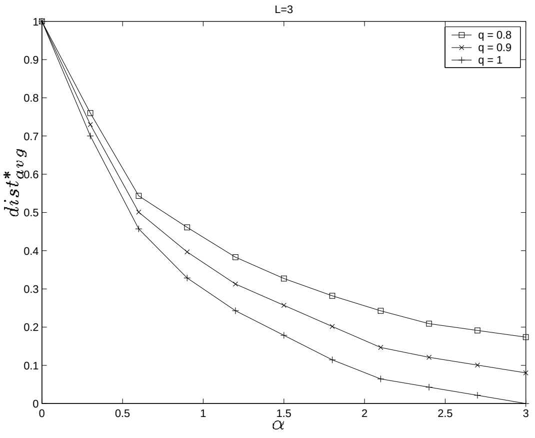


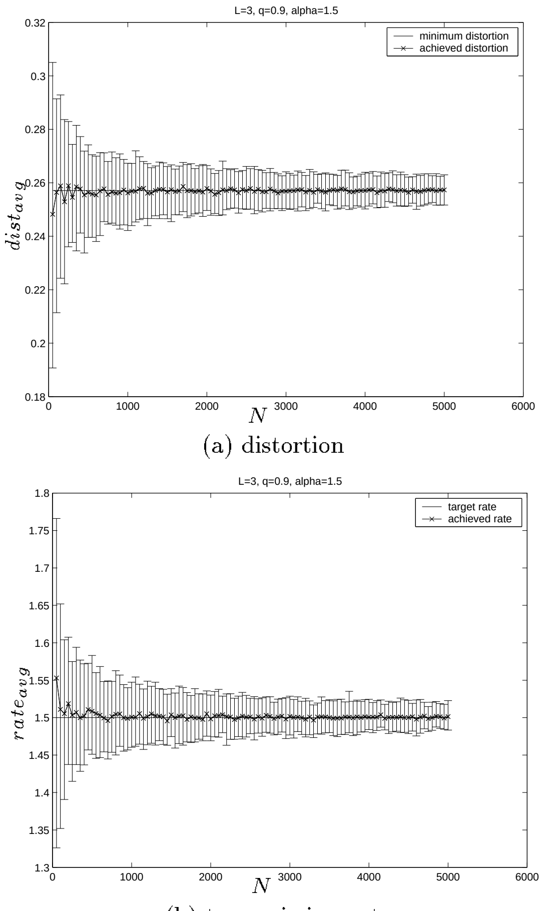
Figure 6: Minimum average distortion for $L = 3$.

ure 2, with the same distortion values. The distortion matrix is given by:

$$[d_{ij}] = \begin{bmatrix} 7 & 3 & 1 & 0 \\ 4 & 3 & 1 & 0 \\ 3 & 2 & 1 & 0 \\ 3 & 1 & 1 & 0 \end{bmatrix} /7. \qquad (13)$$

Figure 6 gives the minimum average distortion $dist^*_{avg}$, as a function of the target transmission rate α, for different values of channel success rate q. We observe that the curve of $dist^*_{avg}$ as a function of α has the shape of an exponential decay: the minimum distortion decreases exponentially with the target transmission rate. It means that high (less conservative) distortions require the transmission of a very small number of packets, while low distortions require to send a very large number of packets.

We study the comparison between optimal dynamic randomized policies, as given by our optimization framework, and some simple static randomized policies. We denote by *simple policy* (l, p), the policy that sends alternatively l layers, with probability p and $l - 1$ layers with probability $1 - p$. Simple policy $(l, 1)$ corresponds to a static non–randomized policy which sends l layers for all frames of the video. Figure 7(a) and Figure 7(b) plot, for different values of q, the minimum average distortion $dist^*_{avg}$ as a function of the target transmission rate α, when using a simple static policy, a Disjoint S+EC optimal dynamic policy and a Joint S+EC optimal dynamic policy. First, we see that, for both channel conditions ($q = 0.95$ and $q = 0.8$), Joint S+EC optimal dynamic policies give the best performance for all values of the target rate α. The increase in distortion when using Disjoint S+EC optimal policies over Joint S+EC optimal policies can go up to 31% and 13%, for $q = 0.95$ and $q = 0.8$ respectively (this corresponds to $\alpha = 0.6$). Second, we observe that Disjoint S+EC optimal policies give similar distortion values as simple static policies. It means that, given our distortion matrix $[d_{ij}]$, optimizing scheduling without considering decoder EC gives similar performance as a very simple scheduling algorithm. This shows that complex optimization procedures for packet scheduling of streaming media can be inadequate when not considering decoder EC. Finally, comparing both figures, note that the difference in performance between all scheduling policies is lower for lower values of q. According to common intuition, control policies at low channel success rates are less efficient than at high success rates.

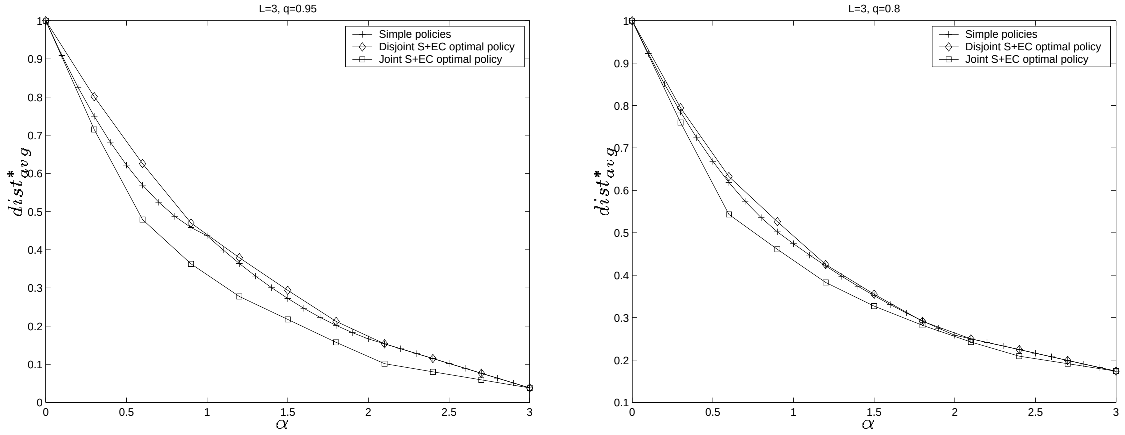


(a) distortion

(b) transmission rate

Figure 8: Simulations with $L = 3$, $q = 0.9$ and $\alpha = 1.5$, for video segments containing up to 5000 frames.

Finally, we show simulations of our Joint S+EC optimal policies for a target transmission rate of $\alpha = 1.5$, over a channel with success rate $q = 0.9$. We averaged our results over 100 channel realizations. Figure 8(a) and Figure 8(b) plot the achieved average distortion and average transmission rate, respectively, as a function of the number of frames of the video N (up to 5000 frames). We plot confidence intervals that represent 95% of the channel runs. As we can see on both figures, as the number of frames increases, the achieved transmission rate and distortion averaged over all channel realizations converge towards the target rate α and the minimum distortion $dist^*_{avg}$, respectively. For a 50 frame segment, the convergence errors are only of 4% for both distortion and transmission rate. However, the confidence intervals are quite large for segments with a low number of frames: for a 50 frame segment, the transmission rate achieved for a given channel realization can be up to 18% higher than α, and the distortion up to 19% higher than $dist^*_{avg}$. For a 500 frame segment, this errors come down to 5% and 7%, respectively.



(a) $q = 0.95$ (b) $q = 0.8$

Figure 7: Comparison between optimal and simple randomized policies for $L = 3$.

Since, in common videos, most homogeneous segments are composed of tens to thousands of frames (homogeneous segments usually correspond to video scenes [9]), we expect that our optimization framework over an infinite horizon will achieve a good operational performance in most cases. For video segments composed of a few frames only, it may be more appropriate to use finite horizon linear programming in order to find optimal policies for each separate frame, as mentioned at the beginning of Section 4.

5. ADDITIONAL QUALITY CONSTRAINT

In Problem 2, we added a new quality constraint to our optimization framework. Specifically, besides minimizing the average distortion, $dist_{avg}$, the optimal transmission policy should also maintain an average variation in distortion between consecutive images, var_{avg}, below a maximum sustainable value γ. As in Problem 1, we consider that the video has infinite length. For a given scheduling policy σ, $var_{avg}(\sigma)$ is the long–run average defined by:

$$var_{avg}(\sigma) := \lim_{n\to\infty} \frac{1}{n-1} E_\sigma[\sum_{i=2}^{n} |D_i - D_{i-1}|] \qquad (14)$$

As for Problem 1, we analyze Problem 2 with a Markov Decision Process over an infinite horizon. The expected average distortion of a given frame n depends only on action A_n and on the state for the previous frame $n-1$, i.e., X_n. However, the expected average variation in distortion for frame n depends also on the value of the state for frame $n-2$, i.e., X_{n-1}. Indeed, from (14), we have $var_{avg}(\sigma) = E_\sigma[|D_n - D_{n-1}|]$, where D_{n-1} is the distortion for frame $n-1$, which depends on the number of layers that have been received for frames $n-1$ and $n-2$, i.e, X_n and X_{n-1} respectively.

We consider the MDP $\{X_{n-1}, X_n, A_n, n = 0, \dots\}$, where $\{X_{n-1}, X_n\}$ and $\{A_n\}$ are the state and action processes. We define the reward and cost functions as:

$$r_n(X_{n-1}, X_n, A_n) = -E[D_n | X_{n-1}, X_n, A_n] \qquad (15)$$

$$c_n(X_{n-1}, X_n, A_n) = A_n, \qquad (16)$$

$$c'_n(X_{n-1}, X_n, A_n) = E[|D_n - D_{n-1}| | X_{n-1}, X_n, A_n]. \qquad (17)$$

From these definitions, Problem 2 can be rewritten as finding an optimal policy σ^* which maximizes the long–run average reward:

$$\lim_{n\to\infty} \frac{1}{n} E_\sigma[\sum_{m=1}^{n} r_m(X_{m-1}, X_m, A_m)]$$
$$\text{s.t.} \begin{cases} \lim_{n\to\infty} \frac{1}{n} E_\sigma[\sum_{m=1}^{n} c_m(X_{m-1}, X_m, A_m)] \le \alpha, \\ \lim_{n\to\infty} \frac{1}{n} E_\sigma[\sum_{m=1}^{n} c'_m(X_{m-1}, X_m, A_m)] \le \gamma, \end{cases} \qquad (18)$$

which falls into the general theory of Markov Decision Processes with multiple constraints. The optimal policy can be found from a linear program similar to the one given for Problem 1. Compared to Problem 1, the number of variables of the LP is increased from $(L+1)^2$ to $(L+1)^3$, and the number of constraints from $L+3$ to $L+4$, which remains tractable. Note that the additional constraint is expressed as follows:

$$c'_n(i, j, a) = (1-q)(\sum_{k=0}^{a-1} |d_{jk} - d_{ij}| q^k) + |d_{ja} - d_{ij}| q^a \qquad (19)$$

Figure 9(a) and Figure 9(b) show, for $L = 3$, the minimum distortion $dist^*_{avg}$ as a function of the target rate α, for selected values of the maximum variation in distortion γ. As we can see on both figures for low values of γ, limiting the variation in distortion comes with an increase in the minimum distortion. We also observe that, for a given channel success rate q and a given target rate α, there exists a minimum value of γ from which the minimum distortion stays constant. Other simulations, which are not shown here, showed that this corresponds to the value of the average variation in distortion achieved by the optimal policy of Problem 1, i.e., without the constraint on var_{avg}.

6. MPEG–4 FGS VIDEOS

Fine Granularity Scalability (FGS) is a new profile of MPEG–4, which has been specifically standardized for transmission of video over the Internet [1]. The FGS–EL can be truncated anywhere before transmission, giving the fine–grained property. There is no motion compensation in the FGS–EL, so that it is highly resilient to transmission errors. According to the MPEG group [2], because the Internet packet loss rate is usually low (under 20%), a typical scenario for transmitting MPEG–4 FGS encoded video over the Internet is to transmit the BL with high reliability (i.e., with channel error correction) and the FGS–EL with no error control. Therefore, we can directly apply our optimization framework to the FGS–EL. We suppose that the BL is transmitted without loss, and that the number of ELs extracted from the FGS–EL is constant for the current video segment (this can be determined by a coarse–grained network–adaptive algorithm, such as in [8, 10]).

In our experiments, we choose the simplest strategy for temporal error concealment, which consists in replacing the missing layers in the current frame by the corresponding layers in the previous frame. During our experiments, we have noticed that this strategy performs well for low motion video segments but very poorly for segments with high motion. Video segments with a fairly high amount of motion, such as *Coastguard* or *Foreman*, would require an error concealment strategy which compensates for motion. For example, [4] presents a scheme for error–concealment in the FGS layer, which uses, together with the layers from the previous frame, the motion information contained in the BL of the current frame. Since we suppose that the BL is transmitted without loss, such a strategy would be easily applicable to our system.

We present experiments with the low motion segment *Akiyo*. We used the Microsoft MPEG–4 software encoder/decoder [7] with FGS functionality. We encode the video into a VBR–BL, with average bitrate of 36 kbps, and a FGS–EL with average bitrate of 900 kbps. We cut the FGS–EL into 3 layers of equal size ($L = 3$). The video segment is encoded into the CIF format (352 × 288 pixels), at a frame rate of 30 frames/sec. It contains $N = 300$ frames. In order to prevent too much fluctuations in quality between successive frames, the first BL frame is encoded as an I–picture and all following frames as P–pictures. (We noticed that our VBR BL–encoder could give important variations in quality between the different types of frames [9], which would be certainly smoothed by a better encoder.)

Figure 10 shows, for different reception states (X_n, X_{n+1}), the PSNR of frame n after error concealment, for n between 100 to 150. Recall that X_{n+1} denotes the number of layers which are available for frame n. We verify that, for a given number of received layers for frame n, the PSNR of frame n increases with X_n. This shows that temporal error concealment is effective in increasing the quality of the rendered video. The increase in quality can be highly significant for some frames. For example, for frame 120, replacing the first FGS–EL of the current frame by the first EL from the previous frame can improve the quality by almost 2 dB (when $X_{121} = 0$, the PSNR of frame 120 goes from 33.2 dB when $X_{120} = 0$ to 35 dB when $X_{120} = 1$).

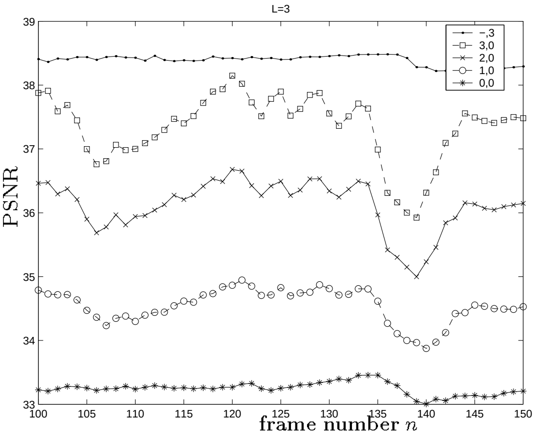


Figure 10: PSNR of frames 100 to 150 of *Akiyo* after EC, for different receiver states (X_n, X_{n+1}).

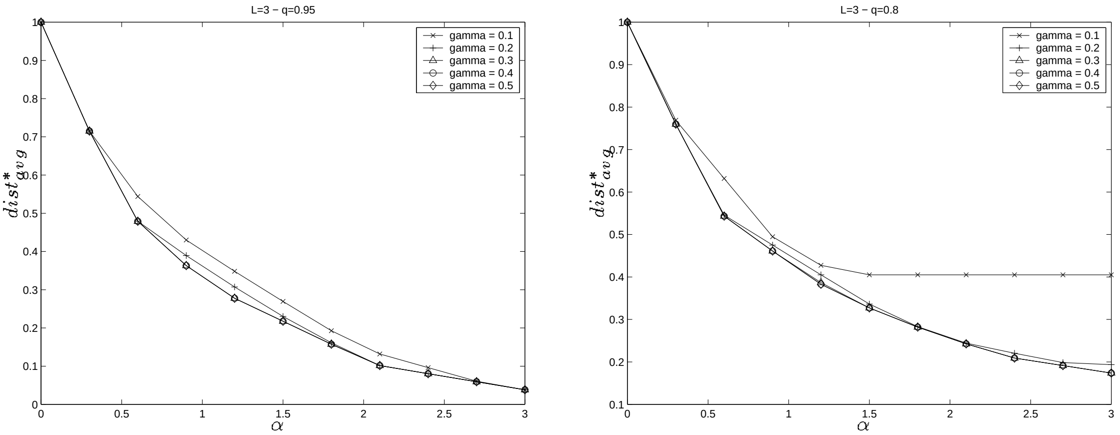
Figure 11 shows a zoomed part of decoded frame 140 after error concealment when no EL was received for frame 140 nor for frame 139 (left), no EL was received for frame 140 but all 3 layers of frame 139 were received (middle), and when all 3 layers of frame 140 were received (right). As we can see, the overall quality of frame 140 is better when all layers of the previous frame have been received (middle picture) than when no layer is available at the receiver for the previous frame (left picture). However, the quality is still lower than when all layers of frame 140 have been received and decoded (right picture).

We computed the average distortion over all frames of the video segment for all possible receiver states. After normalizing, we obtained the following distortion matrix for *Akiyo*:

$$[d_{ij}]_{\text{akiyo}} = \begin{bmatrix} 1 & 0.57 & 0.20 & 0 \\ 0.64 & 0.57 & 0.20 & 0 \\ 0.33 & 0.52 & 0.20 & 0 \\ 0.15 & 0.32 & 0.03 & 0 \end{bmatrix}. \qquad (20)$$

Note that $d_{21} > d_{20}$ and $d_{31} > d_{30}$. This means that replacing all layers from the current frame by layers from the previous frame achieves a lower distortion (better quality) than using the first layer of the current frame and the subsequent layers of the previous frame. This is due to our simple temporal EC strategy. Since we did not implement any motion compensation for EC, the replacement of layers of the current frame by layers of the previous frame create some visual impairments. These impairments are usually minor for low–motion video segments. However, for frames which are significantly different from the previous frames, the resulting increase in distortion can be slightly higher than the decrease in distortion brought by error concealment.

Figure 12(a) and Figure 12(b) show, for Problem 1, the maximum average quality in PSNR as a function of the target rate α, for Joint and Disjoint S+EC optimal policies, as well as simple policies. As in section 4.3, the maximum quality achieved by Disjoint S+EC optimal policies and simple policies is similar, while Joint S+EC optimal policies achieve the best quality for all target rates. The gain brought by jointly optimizing scheduling and error concealment is up to 1.5 dB



(a) $q = 0.95$ (b) $q = 0.8$

Figure 9: Minimum distortion as a function of the target transmission rate α, for $L = 3$.

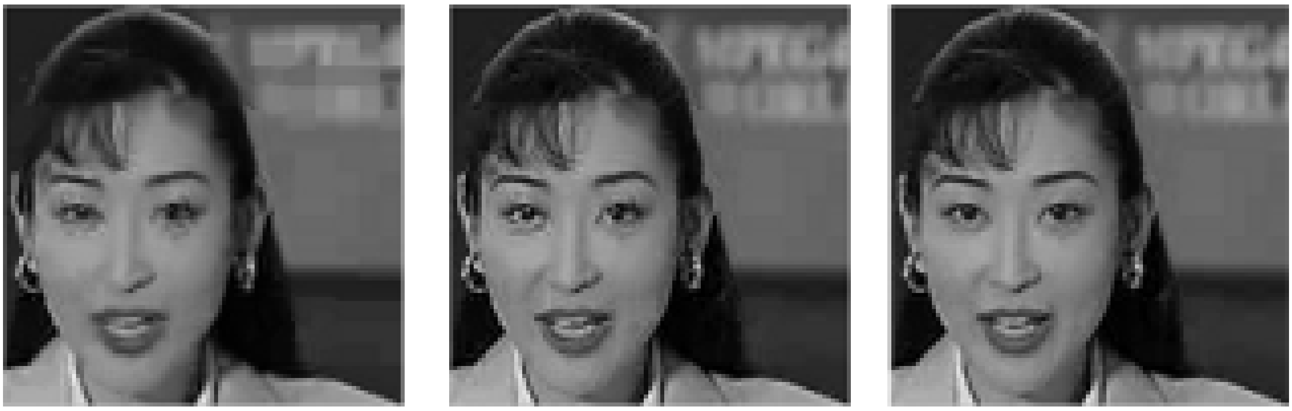
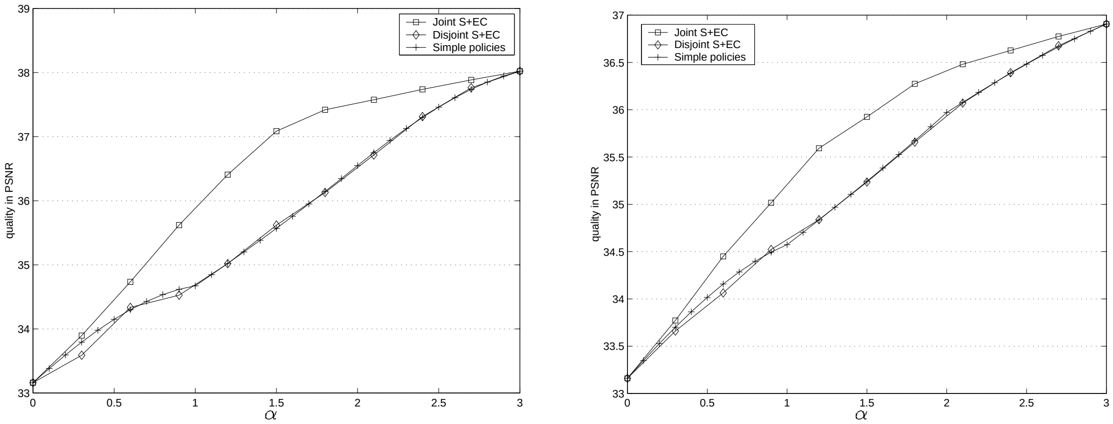


Figure 11: frame 140 of *Akiyo* when (left) $(X_{140} = 0, X_{141} = 0)$ – $PSNR = 33$ dB, (middle) $(X_{140} = 3, X_{141} = 0)$ – $PSNR = 36.3$ dB, (right) $(X_{141} = 3)$ – $PSNR = 38.3$ dB.



(a) $q = 0.95$ (b) $q = 0.8$

Figure 12: Maximum PSNR as a function of the target average transmission rate for *Akiyo*, $L = 3$.

	quality in PSNR			transmission rate		
	target	average	min.	target	average	max.
$\alpha = 1.0$	35.69	35.71	35.48	1.00	1.01	1.05
$\alpha = 1.5$	36.63	36.64	36.36	1.50	1.51	1.62
$\alpha = 2.0$	37.15	37.14	36.93	2.00	2.01	2.12

Table 1: Simulations with $q = 0.9$

for a channel with a high success rate ($q = 0.95$). We expect the difference in quality between Joint and Disjoint S+EC optimal policies to be even higher with error concealment schemes that compensate for motion, notably by using the BL information.

Finally, Table 1 presents the results of simulations over 100 realizations of a packet erasure channel with success rate $q = 0.9$. We show, for different values of the average target transmission rate α, the achieved quality in PSNR and the transmission rate, averaged over all channel realizations. We also show the minimum quality and maximum transmission rate, which are achieved by one channel realization. As we can see, the average achieved values are very close to the target values. This shows that applying our infinite–horizon optimization framework to finite–length videos gives very good performance. In this example, the difference between the minimum achieved PSNR and the target quality is always lower than 0.3 dB. The difference between the maximum achieved transmission rate and the target transmission rate is always lower than 8%.

7. CONCLUSION

In this paper, we have proposed a new optimization framework for joint packet scheduling and error concealment of layered–video (Joint S+EC). We used results on constrained Markov Decision Processes over an infinite horizon, to compute optimal policies with very low–complexity and for a wide range of quality metrics.

We analyzed the problem of minimizing the average distortion under a limited transmission rate. Our analysis leads to a low–complexity algorithm, based on Linear Programming. We showed the potential quality gain brought by Joint S+EC optimization over Disjoint S+EC optimization. We did numerical simulations over a packet–erasure channel with instant feedback, in order to assess the performance of our optimization framework for finite–length videos. We have seen that our method fits particularly well to video segments composed of hundreds of video frames. We showed that our framework allows to accommodate additional quality metrics other than the average distortion, such as the variation in distortion between consecutive images, with no much increase in computational complexity. Finally, we have evaluated the performance of our optimization framework in the context of streaming MPEG–4 FGS videos.

In future work, we plan to extend our framework to a network model with delayed feedback and bursty errors. Also, we will investigate jointly optimizing scheduling and channel error correction, by using FEC codes as additional layers.

8. REFERENCES

[1] *ISO/IEC JTC1/SC29/WG11 Information Technology — Generic Coding of Audio-Visual Objects : Visual ISO/IEC 14496-2 / Amd X*, December 1999.

[2] *ISO/IEC JTC1/SC29/WG11 N4791 — Report on MPEG-4 Visual Fine Granularity Scalability Tools Verification Tests*, May 2002.

[3] E. Altman. *Constrained Markov Decision Processes.* Chapman and Hall, 1999.

[4] H. Cai, G. Shen, F. Wu, S. Li, and B. Zeng. Error Concealment for Fine Granularity Scalable Video Transmission. In *Proc. of IEEE ICME*, pages 145–148, Lausanne, Switzerland, September 2002.

[5] P. A. Chou and Z. Miao. Rate–Distortion Optimized Sender–Driven Streaming over Best–Effort Networks. In *Workshop on Multimedia Signal Processing*, pages 587–592, October 2001.

[6] P. A. Chou and Z. Miao. Rate–Distortion Optimized Streaming of Packetized Media. *submitted to IEEE Transactions on Multimedia*, February 2001.

[7] Microsoft Corp. ISO/IEC 14496 Video Reference Software. Microsoft–FDAM1–2.3–001213.

[8] P. de Cuetos, P. Guillotel, K. W. Ross, and D. Thoreau. Implementation of Adaptive Streaming of Stored MPEG–4 FGS Video. In *Proc. of IEEE ICME*, pages 405–408, Lausanne, Switzerland, August 2002.

[9] P. de Cuetos, M. Reisslein, and K. W. Ross. Streaming FGS–Encoded Video: Insights from a Large Library of Rate–Distortion Traces. Submitted (http://www.eurecom.fr/˜decuetos), December 2002.

[10] P. de Cuetos and K. W. Ross. Adaptive Rate Control for Streaming Stored Fine-Grained Scalable Video. In *Proc. of NOSSDAV*, pages 3–12, Miami, Florida, May 2002.

[11] C. Derman. *Finite State Markovian Decision Processes.* Academic Press, New York, 1970.

[12] U. Horn, K. Stuhlmuller, M. Link, and B. Girod. Robust Internet Video Transmission Based on Scalable Coding and Unequal Error Protection. *Signal Processing: Image Communication*, 15:77–94, 1999.

[13] L. C. M. Kallenberg. *Linear Programming and Finite Markovian Control Problems.* Mathematisch Centrum, Amsterdam, 1983.

[14] N. Karmarkar. A New Polynomial Time Algorithm for Linear Programming. *Combinatorica*, (4):373–395, 1984.

[15] Z. Miao and A. Ortega. Expected Run–time Distortion Based Scheduling for Delivery of Scalable Media. In *Proc. of International Conference of Packet Video*, Pittsburg, PA, April 2002.

[16] M. Podolsky, M. Vetterli, and S. McCanne. Limited Retransmission of Real-Time Layered Multimedia. In *IEEE Workshop on Multimedia Signal Processing*, pages 591–596, Los Angeles CA, December 1998.

[17] R. Rejaie, D. Estrin, and M. Handley. Quality Adaptation for Congestion Controlled Video Playback over the Internet. In *Proc. of ACM SIGCOMM*, pages 189–200, Cambridge, September 1999.

[18] R. Rejaie and A. Reibman. Design Issues for Layered Quality-Adaptive Internet Video Playback. In *Proc. of the Workshop on Digital Communications*, pages 433–451, Taormina, Italy, September 2001.

[19] K. W. Ross. Randomized and Past–Dependent Policies for Markov Decision Processes With Multiple Constraints. *Operations Research*, 37(3):474–477, May–June 1989.

[20] S. D. Servetto. *Compression and Reliable Transmission of Digital Image and Video Signals.* PhD thesis, University of Illinois, May 1999.

[21] Y. Wang and Q. Zhu. Error Control and Concealment for Video Communications: A Review. *Proc. of the IEEE*, 86(5):974–997, May 1998.

[22] Q. Zhang, W. Zhu, and Y-Q. Zhang. Resource Allocation for Multimedia Streaming over the Internet. *IEEE Transactions on Multimedia*, 3(3):339–335, September 2001.