Quality metrics for measuring the end-to-end distortion in packet-switched video communication systems

Yiftach Eisenberg, Fan Zhai, Thrasyvoulos N. Pappas, Randall Berry, Aggelos K. Katsaggelos
Northwestern University, ECE Department, 2145 Sheridan, Evanston, IL, 60208
Email: (yeisenbe, fzhai, pappas, rberry, aggk)@ece.northwestern.edu

ABSTRACT

A critical component of any video transmission system is an objective metric for evaluating the quality of the video signal as it is seen by the end-user. In packet-based communication systems, such as a wireless channel or the Internet, the quality of the received signal is affected by both signal compression and packet losses. Due to the probabilistic nature of the channel, the distortion in the reconstructed signal is a random variable. In addition, the quality of the reconstructed signal depends on the error concealment strategy. A common approach is to use the expected mean squared error of the end-to-end distortion as the performance metric. It can be shown that this approach leads to unpredictable perceptual artifacts. A better approach is to account for both the mean and the variance of the end-to-end distortion. We explore the perceptual benefits of this approach. By accounting for the variance of the distortion, the difference between the transmitted and the reconstructed signal can be decreased without a significant increase in the expected value of the distortion. Our experimental results indicate that for low to moderate probability of loss, the proposed approach offers significant advantages over strictly minimizing the expected distortion. We demonstrate that controlling the variance of the distortion limits perceptually annoying artifacts such as persistent errors.

Keywords: Video quality metrics, end-to-end distortion, efficient resource allocation

1. INTRODUCTION

Network-based video applications have dramatically increased in popularity in recent years. One of the major challenges in supporting these applications is that packet delivery is not guaranteed in transmission systems, such as wireless channels and the Internet. Thus, from the point-of-view of the transmitter, the distortion at the receiver is a random variable that depends on the probability of packet loss in the channel. To illustrate this point, consider the two reconstructed frames shown in Fig. 1. These frames correspond to what would be seen at the receiver given two different channel loss realizations[1]. Clearly, the distortion at the receiver depends on which packets are lost. Thus, to efficiently encode and transmit a video sequence, the metric used to evaluate the end-to-end distortion, D_{tot}, must account for the probabilistic nature of the channel.

(a) (b) (c)

Fig. 1. (a) Expected reconstructed frame averaged over all possible loss simulations. (b and c) Reconstructed frames for two different channel loss realizations.

[1] Note that the same probability of packet loss is used in both simulations.

Recently, there has been considerable research in the area of distortion estimation and characterization for packet-based video transmission systems [1],[2],[6-20],[27]. One of the most common approaches is to use the expected distortion between the original and the reconstructed sequence at the receiver as the performance metric. In this paper, we explore the benefits of accounting for the variance, as well as the mean, of the end-to-end distortion when allocating limited source and channel resources. By accounting for the variance of the distortion, the proposed approach makes it more likely that what the end-user sees closely resembles what was transmitted, thus increasing the reliability of the system. This paper builds on our prior work, some of which can be found in [1], [2], and [3].

At the receiver, the end user sees only one of many possible reconstructed sequences, depending on the actual realization of packet losses. Therefore, the actual distortion at the receiver does not necessarily equal the expected distortion. To illustrate this point, consider the images shown in Fig. 1. While the expected reconstructed frame (averaged over all possible loss realizations) may be reasonable, as shown in Fig. 1(a), the quality at the receiver may vary greatly based on which packets are lost, as shown in Fig. 1 (b) and (c). Therefore, we argue that the variance of the end-to-end distortion should also be considered when characterizing video quality in lossy packet networks. We develop the concept of "Variance-Aware per-Pixel Optimal Resource-allocation" (VAPOR), and present a framework for controlling both the expected value and the variance of the end-to-end distortion. In Sect. 6, experimental results demonstrate that reducing the variance of the distortion can help limit perceptually annoying artifacts such as error propagation. Video conferencing, target tracking, surveillance, and personal communications are just a few applications that can benefit from variance-aware resource allocation.

The remainder of the paper is organized as follows. Next we describe the system model. In Sect. 3, we characterize the end-to-end distortion in packet-based video communication systems. Two variance-aware resource allocation formulations are presented in Sect. 4, followed by a discussion of the solution approach in Sect. 5. Extensive experimental analysis is presented in Sect. 6. Section 7 contains concluding remarks.

2. SYSTEM MODEL

We begin by providing a high-level overview of a packet-based video transmission system. Figure 2 highlights some of the major conceptual components in such a system. The original video signal is first compressed by the *video encoder*. Compression reduces the number of bits used to describe the video sequence by exploiting both temporal and spatial redundancy. The encoded video will be transmitted over a communication channel that is lossy by nature. Therefore, the video sequence must be encoded in an error resilient way that minimizes the effects of losses on the decoded video quality. A recent review of resilient video coding techniques can be found in [4].

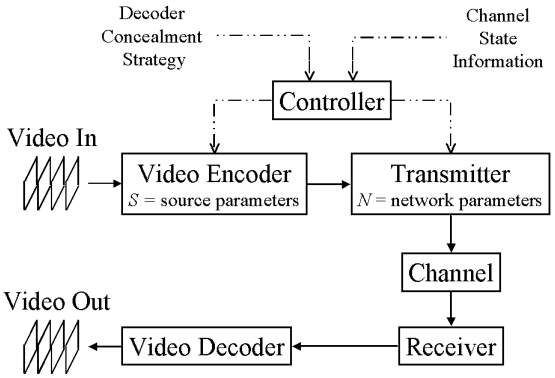


Fig. 2. Packet-based video transmission system diagram.

In this paper, we focus on one of the most widely utilized video coding techniques, Block-based Motion Compensated (BMC) video coding (e.g., H.263 and MPEG-4). In this approach, each frame is divided into macro-blocks (MBs) that can either be independently encoded (Intra coded) or predictively coded from a reference MB in a previous frame (Inter coded). For Inter coding, a motion vector (MV) specifies the location of the reference MB in the reference frame. Hence the name "motion compensated" video coding. Temporal prediction offers increased coding

efficiency over Intra coding but is susceptible to error propagation. Transform coding followed by quantization and entropy coding complete the BMC coding process. Let S denote the source coding parameters, such as the coding mode and quantization step-size for all the MBs in a frame (or group of frames). The selection of S affects the source bit rate as well as the end-to-end distortion.

The encoded video sequence is then transmitted over a communication channel. This typically involves packetizing the video stream and passing the packets through the appropriate protocol layers (e.g., RTP/UDP/IP). In Fig. 2, this functionality is implemented in the *transmitter* block. We take a high level view of this block in order to emphasize similarities between a wide range of applications, such as wireless and Internet-based video transmission. Let N represent the set of *network parameters* that can be controlled at the transmitter. This set of parameters depends on the application. For example, in wireless communications, the transmission power and rate can be adapted; in a Differentiated Services (DiffServ) network, the priority (or QoS) assigned to each packet can be considered a network parameter. In most communication systems, some form of *channel state information* (CSI) is available at the sender, such as an estimate of the fading level in a wireless channel or the congestion over a route in the Internet. Based on the channel state information θ and the choice of network parameters N, the transmitter is able to estimate channel characteristics, such as the probability of packet loss ρ. We indicate this relationship as $\rho = f(N,\theta)$, where the function f can be determined analytically or from empirical data.

At the *receiver*, the demodulated bit stream is processed by the channel decoder, which performs error detection and/or correction. Corrupt packets can either be passed onto the video decoder or discarded. Here, we assume that only error free information is passed to the video decoder and that corrupt packets are considered lost. This is motivated by the fact that in Internet-based communications, the probability of packet corruption is very low compared to the probability of packet loss caused by congestion in the network. Similarly, in wireless communications, the probability of an error being undetected is far smaller than the likelihood of a packet being lost due to a deep fade in the channel.

The *video decoder* is responsible for reconstructing the video sequence for display at the receiver. Because some encoded information may have been lost, e.g., due to buffer overflow at a router, the video decoder must conceal any lost information. The commonality among all error concealment strategies is that they exploit redundancy in the received video sequence to conceal lost information. One popular concealment strategy, which we employ in our experimental results, is to use temporal replacement based on the motion information of neighboring macro-blocks. A comprehensive review of error concealment techniques can be found in [5].

The *controller* block in Fig. 2 indicates the component of the video transmission system responsible for adapting the source coding parameters S and the network parameters N based on knowledge of the concealment strategy, the source content and any available CSI. The selection of S and N affects the end-to-end distortion D_{tot}, the end-to-end delay T_{tot}, and the total cost C_{tot} for delivering the video sequence to the end-user. We will use $D_{tot}(S,N)$, $C_{tot}(S,N)$, and $T_{tot}(S,N)$ to explicitly indicate these dependencies. Distortion is caused by both source coding artifacts and channel errors, and will be discussed in greater detail in Sect. 3. The cost C_{tot} is a measure of the limited resources consumed in transmitting the video sequence, and will be discussed more in Sect. 4. The end-to-end delay T_{tot} is the time between when a video frame is captured at the transmitter and when it is displayed at the receiver. T_{tot} depends in part on the number of bits used to encode the sequence, the transmission rate and any scheduling decisions made by the transmitter.

3. END-TO-END DISTORTION METRICS

A critical component in any video communication system is a metric for evaluating the quality of the received video signal. Recently, there has been considerable research in the area of distortion characterization and estimation for packet-based video transmission systems. Here, we highlight two general classes of distortion metrics. The first class consists of methods that measure video quality by the expected distortion in the received sequence. The second consists of metrics whose aim is to produce more smooth quality by accounting for several sources of distortion variation.

In this section we present a framework for characterizing the distortion between the original video and the reconstructed sequence at the receiver. This framework is based on knowledge of how the original sequence is encoded, the probability of packet loss in the channel, and the concealment strategy used by the decoder. Since a pixel is the smallest information symbol in a digital video sequence, we use per-pixel accurate expressions to characterize the end-to-end distortion.

Consider a single pixel in the video sequence. For the system described in Sect. 2, we assume the encoded information for each pixel is contained in only one packet and that each packet is either received correctly or lost. In this

case, the distortion between the original and reconstructed value for the *i*th pixel of the *k*th packet in the *n*th frame is a random variable with the following distribution

$$D^{n,k,i} = \begin{cases} D_R^{n,k,i} & \text{with probability } (1-\rho^{n,k}) \\ D_L^{n,k,i} & \text{with probability } (\rho^{n,k}) \end{cases}, \quad (1)$$

where $\rho^{n,k}$ is the probability that the *k*th packet is lost, $D_R^{n,k,i}$ is the distortion if the packet is *received* correctly, and $D_L^{n,k,i}$ is the distortion if the packet is *lost*. If the pixel is predictively encoded, then $D_R^{n,k,i}$ is a random variable due to random losses in previous frames. On the other hand, if the pixel is independently encoded (Intra coded), then $D_R^{n,k,i}$ is deterministic. The distortion if a packet is lost depends on the concealment strategy used at the decoder. If prediction is used in the concealment strategy, e.g., temporal concealment, then $D_L^{n,k,i}$ is also a random variable.

3.1. Expected distortion metrics

The expected value of the end-to-end distortion for a given pixel can be written as

$$E[D^{n,k,i}] = (1-\rho^{n,k})E[D_R^{n,k,i}] + (\rho^{n,k})E[D_L^{n,k,i}], \quad (2)$$

where $E[\bullet]$ indicates the expected value taken with respect to the probability of loss. The average expected distortion for the *k*th packet in the *n*th frame is defined as $\overline{D}^{n,k} = (1/I^k)\sum_{i=1}^{I^k} E[D^{n,k,i}]$, where I^k is the number of pixels in the *k*th packet. Similarly, the average expected distortion for the *n*th frame and for the sequence are simply $\overline{D}^n = (1/K)\sum_{k=1}^{K} \overline{D}^{n,k}$ and $\overline{D}_{seq} = (1/M)\sum_{n=1}^{M} \overline{D}^n$, respectively, where *M* is the number of frames in the sequence.

Work on resilient video coding and transmission for packet lossy networks has primarily focused on minimizing the average expected distortion for a frame (or group of frames) [6-20],[27]. In these approaches the end-to-end distortion is defined as $D_{tot}^n = \overline{D}^n$. The average expected distortion is somewhat of a natural measure of the end-to-end distortion because it accounts for the source coding distortion as well as the expected channel induced distortion. As discussed in the following sections, the drawback of defining the end-to-end video quality as $\overline{D}^n$, is that it does not capture undesirable perceptual artifacts such as large variations in quality. Several methods have been proposed for calculating the average expected distortion. The most straightforward approach is to simulate several packet loss patterns at the encoder and to average the resulting distortions [7]. Although this approach produces a better estimate of the expected distortion as the number of simulations increases, the drawback is that the computational complexity and storage requirements can quickly become impractical.

Methods for accurately calculating the expected distortion have recently been proposed [8], [9], [10]. The main contribution of these approaches is that they show that under certain conditions, it is possible to accurately compute the expected distortion with finite storage and computational complexity by using per-pixel accurate recursive calculations. In [8], Hind's method is based on recursively calculating the distribution of the reconstructed value for each pixel in a frame. A recursively accurate method for calculating the expected mean absolute difference is presented in [10]. In [9], Zhang et al. develop a powerful algorithm called ROPE, which efficiently calculates the expected mean squared error by recursively computing only the first and second moment of each pixel in a frame. In many advanced video coding schemes, e.g., H.26L and MPEG-4, non-integer motion compensation, deblocking filters, and complex concealment strategies introduce cross-correlation between pixels that make ROPE less precise. Recently there has been work on approximating these cross-correlation terms in order to extend ROPE to more sophisticated coding schemes [11], [12].

Model-based distortion estimation methods have also been proposed and are useful when the computational complexity and storage capacity are limited. In [13], the authors present a recursive distortion estimation algorithm, which only differs slightly from ROPE in that they approximate the distortion due to concealment. In order to estimate the expected distortion, a likely subset of the possible loss patterns is considered in [14]. In [15], He et al. develop a model for estimating both source and channel distortion based on the Intra refresh rate and the percentage of zeros among the quantized transform coefficients. Another popular metric for calculating the expected distortion is to

consider the reduction in distortion given that a packet and all the packets it depends on are received, as in [16]. This approach works well when the dependencies between packets are clearly defined, e.g., in progressive and scalable coding, but does not explicitly take into account error propagation due to concealment.

3.2. Variance-aware distortion metrics

In video coding and transmission there are many sources of quality variation. We review several variance-aware distortion metrics whose aim is to reduce these variations in order to smooth out the quality of the received video sequence. Reducing the spatial variation in quality across a frame has been considered in order to prevent having some regions of a frame with good quality and others with relatively poor quality. In [17] and [18], one attempt at producing more even quality was to minimize the maximum distortion within a frame, as opposed to the average distortion, i.e., $D_{tot}^{n} = \max_{k}\{E[\overline{D}^{n,k}]\}$.

Controlling temporal variations in quality has also been considered. Rate-control, i.e., assigning bandwidth (bits) to the different frames in a sequence, is related to this type of variation [19]. By allocating more bandwidth during periods of high activity and less to frames with little motion, a rate-control scheme can reduce the overall distortion within a specified time window. Similarly, approaches such as [18], [20], have looked at the benefits of limiting large temporal variations in distortion across a group of frames.

In video transmission applications, a third source of quality variation is the variance in distortion caused by channel errors. The remainder of this paper addresses this source of quality variation. While the expected distortion (averaged over all possible loss realizations) may be reasonable, the quality at the receiver may vary greatly based on which packets are lost. Therefore, to determine the likelihood that the actual distortion seen by the end-user is near the expected distortion, one must look at the variance of the distortion.

The variance of the distortion for a given pixel is by definition equal to $Var[D^{n,k,i}] = E[(D^{n,k,i})^2] - E[D^{n,k,i}]^2$. By substituting (1) into the previous equation and rearranging terms, we can express $Var[D^{n,k,i}]$ as

$$Var[D^{n,k,i}] = (1-\rho^{n,k})Var[D_R^{n,k,i}] + (\rho^{n,k})Var[D_L^{n,k,i}] + (1-\rho^{n,k})(\rho^{n,k})\left\{E[D_R^{n,k,i}] - E[D_L^{n,k,i}]\right\}^2, \quad (3)$$

where $Var[D_R^{n,k,i}]$ and $Var[D_L^{n,k,i}]$ are the variance in distortion if the packet containing the coding information for this pixel is received and lost, respectively. As expected, $Var[D^{n,k,i}]$ increases when $Var[D_R^{n,k,i}]$ or $Var[D_L^{n,k,i}]$ increase. Therefore, Intra coding, which has $Var[D_R^{n,k,i}] = 0$, enables the transmitter to greatly decrease $Var[D^{n,k,i}]$. Inter coding on the other hand is susceptible to temporal error propagation and thus has $Var[D_R^{n,k,i}] \geq 0$. This suggests that $Var[D_R^{n,k,i}]$ is a good indicator of how severely a packet may be affected by error propagation. From (3), we see that $Var[D^{n,k,i}]$ increases as $\left\{E[D_R^{n,k,i}] - E[D_L^{n,k,i}]\right\}^2$ increases. This means that if a pixel is difficult to conceal, i.e., $E[D_L^{n,k,i}] >> E[D_R^{n,k,i}]$, its distortion varies greatly depending on whether the packet is received or lost.

From (3), it can be seen that $Var[D^{n,k,i}]$ is a concave quadratic function of $\rho^{n,k}$. This is intuitively satisfying since there is less variability in $D^{n,k,i}$ when the probability of loss is either very small or very large. For example, when $\rho = 0$ or 1 for all the packets in the sequence, $Var[D^{n,k,i}] = 0$.

Let $Std[D^{n,k,i}] = \sqrt{Var[D^{n,k,i}]}$ represent the standard deviation of the distortion for a given pixel. The average standard deviation in distortion for the *k*th packet in the *n*th frame is defined as $\overline{\sigma}^{n,k} = (1/I^k)\sum_{i=1}^{I^k} Std[D^{n,k,i}]$. Similarly, the average standard deviation in distortion for the *n*th frame and for the sequence are $\overline{\sigma}^{n} = (1/K)\sum_{k=1}^{K}\overline{\sigma}^{n,k}$ and $\overline{\sigma}_{seq} = (1/M)\sum_{n=1}^{M}\overline{\sigma}^{n}$, respectively.

It is important to note that the average variance in distortion per pixel is not equal to the variance of the average distortion for a frame, i.e.,

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{I^k}\sum_{i=1}^{I^k}Var[D^{n,k,i}] \neq Var\left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{I^k}\sum_{i=1}^{I^k}D^{n,k,i}\right]. \quad (4)$$

The expression on the right hand side is more difficult to compute since it requires calculating the cross-correlation between the distortion of all the pixels in the frame. In addition, the problem of optimally encoding and transmitting a video frame based on the variance in quality, as done in Sect. IV, becomes infeasible if the $Var\left[(1/K)\sum_{k=1}^{K}\frac{1}{I^k}\sum_{i=1}^{I^k}D^{n,k,i}\right]$ is used. Another motivation for using $(1/K)\sum_{k=1}^{K}\frac{1}{I^k}\sum_{i=1}^{I^k}Var[D^{n,k,i}]$ is that it captures local variations in quality better than the variance of the average distortion.

3.3. Recursive calculations based on the Mean Squared Error

The expressions for the mean (2) and variance (3) of the end-to-end distortion derived in the previous sections hold for a wide range of distortion metrics. In our experimental results, we consider the case where distortion is defined as the squared error between the original pixel value x and the reconstructed pixel value at the receiver $\tilde{x}$, i.e., $D=(x-\tilde{x})^2$. In this case, the first two moments of the reconstructed pixel value, i.e., $E[\tilde{x}]$ and $E[(\tilde{x})^2]$, are needed to accurately calculate $E[D]=E[(x-\tilde{x})^2]$. Similarly, the first four moments of the reconstructed pixel value are needed to accurately calculate $Var[D]=E[(x-\tilde{x})^4]-E[(x-\tilde{x})^2]^2$.

In certain cases, optimal distortion estimation methods, such as ROPE [9] and [8] can be used to accurately and recursively calculate the necessary reconstructed pixel moments. In other cases, such as non-integer pixel motion compensation, models may be needed to estimate the expected distortion. Developing efficient models for estimating the variance of the end-to-end distortion is an area requiring future research.

4. VARIANCE-AWARE RESOURCE ALLOCATION

Our goal is to control both the source-coding and transmission parameters (S, N) in order to minimize the end-to-end distortion while using a limited amount of transmission cost and delay. We can formally write this optimization as

$$\min_{\{S,N\}} D_{tot}^{n}(S,\boldsymbol{N}) \text{ subject to: } C_{tot}^{n}(S,N)\le C_0^{n} \text{ and } T_{tot}^{n}(S,N)\le T_0^{n}, \quad (5)$$

where D_{tot}^{n} is the end-to-end distortion, C_{tot}^{n} is the total transmission cost, C_0^{n} is the transmission cost constraint, T_{tot}^{n} is the total transmission delay, and T_0^{n} is the maximum transmission delay for the nth frame. In (5), the transmission cost C_{tot}^{n} corresponds to the limited resources consumed in the transmission of the nth frame. For example, in wireless communications, mobile users typically rely on a limited battery supply. Therefore, in wireless video communications, C_{tot}^{n} represents the energy consumed transmitting the nth frame. In a Differentiated Services (DiffServ) Internet, C_{tot}^{n} may represent the price paid to transmit packets at different quality of service. We assume that a higher-level controller assigns cost and delay constraints (C_0^{n} and T_0^{n}) per frame based on the application. The design of this controller is outside the scope of this paper.

In the previous sections, we discussed the need to account for both the mean and the variance of the end-to-end distortion when evaluating video quality. One way to do this is by defining the distortion for a given frame in (5) as the weighted sum of the expected distortion plus the standard deviation in distortion, i.e.,

$$D_{tot}^{n}=(1-\alpha)\overline{D}^{n}+(\alpha)\overline{\sigma}^{n}=\frac{1}{I}\sum_{k=1}^{K}\sum_{i=1}^{I^k}(1-\alpha)E\left[D^{n,k,i}\right]+(\alpha)Std\left[D^{n,k,i}\right], \quad (6)$$

where $\alpha \in [0,1]$ defines the relative importance of the variance of the end-to-end distortion. We use $Std[D^{n,k,i}]$ instead of $Var[D^{n,k,i}]$ so that the units of D_{tot}^{n} are consistent. By using (6) as the objective in the optimization problem (5) and solving for different values of α, we can observe the trade-off between minimizing the expected value and the variance of the distortion. For example, in Fig. 3, we plot the mean and the standard deviation in distortion versus α for frame 43 of the "foreman" test sequence. As shown in Fig. 3, increasing α reduces the standard deviation at the cost of increased expected distortion, as expected. When $\alpha = 0$, we obtain a special case of the general formulation in which the objective is to strictly minimize the expected distortion per frame, as in [6-20]. As shown in Fig. 3, it is possible to significantly reduce $\overline{\sigma}^{n}$ while only slightly increasing $\overline{D}^{n}$. This observation motivates the following alternative variance-aware formulation.

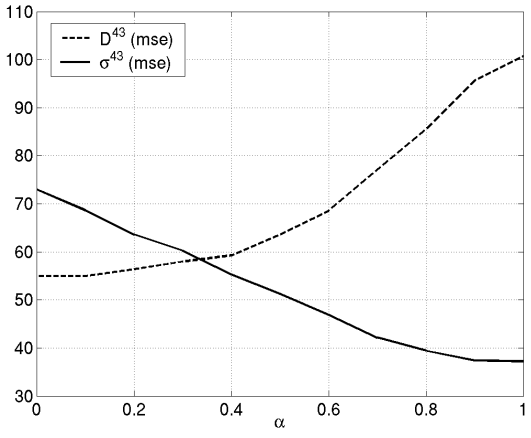


Fig. 3. Impact of α on the average expected distortion (D^{43}) and average standard deviation in distortion (σ^{43}) for frame 43 of the "foreman" sequence coded at 30 fps with R = 150Kbps and ρ = 0.01.

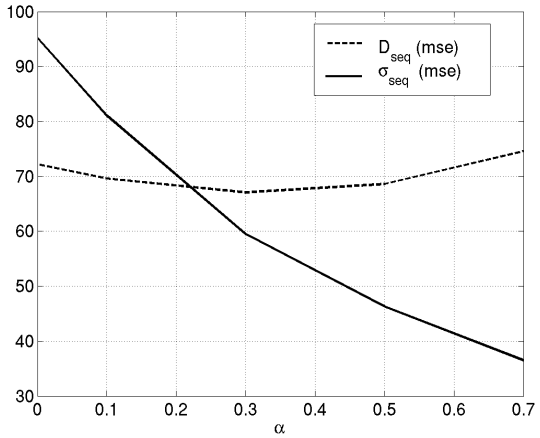


Fig. 4. The impact of α on the expected distortion (D_{seq}) and standard deviation in distortion (σ_{seq}) averaged over all the frames of the "foreman" sequence coded at 30 fps with R = 150Kbps and ρ = 0.01.

By adjusting α, we can control the value of both the mean and the variance of the distortion. Let $\overline{D}_{\min}^{n}$ be the minimum expected distortion for the frame (i.e., the solution to (5) with α=0). We modify (5) slightly to include the optimal selection of α, by constraining the difference between the expected distortion $\overline{D}^{n}$ and the minimum achievable value $\overline{D}_{\min}^{n}$. In other words, we solve the following optimization

$$\min_{\{S,N,\alpha\}} D_{tot} = (1-\alpha)\overline{D}^{n} + (\alpha)\overline{\sigma}^{n} \quad \text{subject to: } C_{tot}^{n}(S,N) \le C_0^{n},\ T_{tot}^{n}(S,N) \le T_0^{n},\ \text{and } \overline{D}^{n} \le \Delta_d \times \overline{D}_{\min}^{n} \tag{7}$$

where $\Delta_d \ge 1$ represents the maximum increase in expected distortion we are willing to tolerate in order to decrease the variance of the distortion. Note that the formulation in (7) is equivalent to minimizing the variance given an expected distortion constraint. In Sect. 6, we show that for a small increase in expected distortion, the standard deviation of the distortion can be reduced significantly.

To the best of our knowledge, the formulations in (5) and (7) are the first to account for both the mean and the variance of the end-to-end distortion. We refer to formulations of this type as a "Variance-Aware per-Pixel Optimal Resource-allocation" (VAPOR) techniques. The formulation in (5) is less complex than (7) in that α is fixed. The drawback is that it may not be very intuitive how to set α, and the same α may not be desirable for every frame. This problem is addressed in (7) because α is optimally set based on a specified tolerable increase in expected distortion.

5. SOLUTION APPROACH

In order to solve (5) we use Lagrange relaxation and Dynamic Programming (DP). First we introduce two Lagrange multipliers, $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$, and solve the relaxed problem

$$\min_{\{S,N\}} D_{tot}^n + \lambda_1 C_{tot}^n + \lambda_2 T_{tot}^n . \quad (8)$$

By appropriately choosing λ_1 and λ_2, the solution to (5) can be obtained within a convex-hull approximation by solving (8) [21], [22]. Various methods, such as cutting-plane or sub-gradient methods, can be used to search for λ_1 and λ_2 [23]. In our experimental results, we use an efficient method developed in [24] that exploits the structure of the problem in (5).

For each choice of λ_1 and λ_2, we can solve (8) using DP. The concealment strategy used at the receiver may introduce dependencies between packets. For example, temporal concealment based on the motion vectors of neighboring packets causes the distortion for a given packet to depend on how its neighboring packet(s) are encoded as well as their probability of loss. DP can be used to efficiently find the optimal source coding and network parameters for each packet in the frame when the dependencies between packets are limited, e.g., to a small neighborhood. For more details please see [22], [21], and [3]. To solve (7), we can iteratively adjust α and solve (5) until the resulting expected distortion $\overline{D}^n$ is less than or equal to $\Delta_d \times \overline{D}_{\min}^n$.

6. EXPERIMENTAL RESULTS

The basic ideas developed throughout this paper are relevant to a wide range of applications. In this section, we focus on one example, i.e., real-time wireless video communications. Extensive experimental results highlight the potential benefits of accounting for both the expected value and the variance of the distortion when allocating limited resources in packet-based video transmission systems. As a comparison to the proposed VAPOR approach, we consider a more traditional approach whose goal is to minimize the expected distortion per frame, as in [6-20]. We refer to this scheme as the Minimum Expected Distortion (MED) approach. As mentioned in Sect. 4, the MED approach is a special case of (5) in which $\alpha = 0$. In Sect. 6.2., we consider the problem of optimal source coding, i.e., we assume that the probability of packet loss can not be controlled. In Sect. 6.3., we present results suggesting that VAPOR offers improved perceptual quality by reducing error propagation.

6.1. Experimental setup

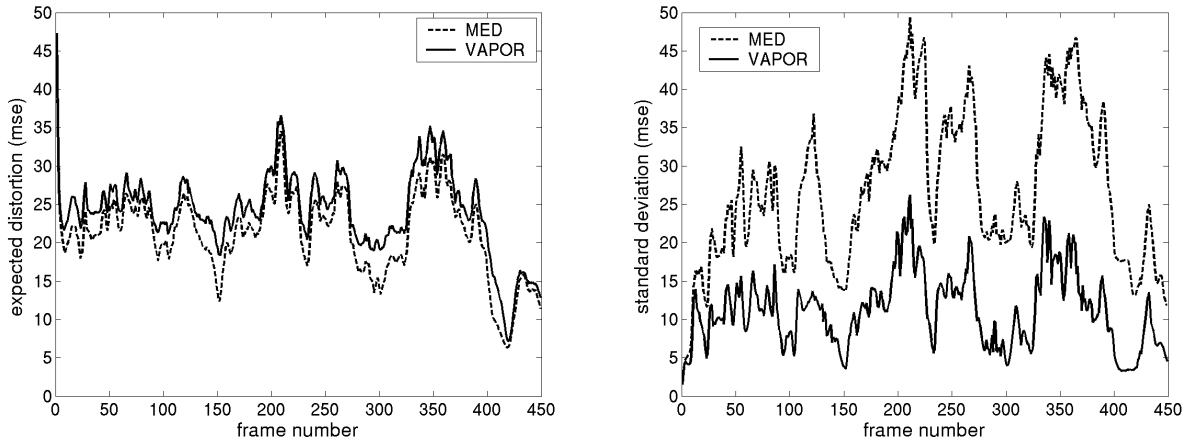
An H.263+ codec is used to encode the video sequences using a limited number of quantization step sizes for "Intra" and "Inter" MBs. In addition, integer pixel motion compensation is used in order to ensure accurate distortion estimation at the transmitter [9]. As in [3], we consider the case where each packet contains a single MB, i.e., each MB is independently decodable. This packetization scheme has low coding efficiency but high resiliency to channel errors.

Similarly, we consider a relatively simple concealment strategy in which the concealment motion vector (MV) for a lost MB is defined as the MV of the MB to the left of the lost MB. If the lost MB is on the left edge of the frame, or if the MB to the left is also lost, then the concealment MV is set to zero. We consider a real-time application with an allowable transmission delay of one frame duration. Therefore, a sequence coded at 30 frames per second (fps) has a delay constraint T_0 = 33 msec. At 15 fps, T_0 = 66 msec. In all the experiments, the "generalized skip mode", introduced in [3], enables the transmitter to intentionally not transmit certain packets if their concealment at the decoder results in adequate quality.

We use the channel model from [3], where each packet is sent over a narrow-band slowly fading channel with additive white Gaussian noise. We model the probability of packet loss ρ in the capacity versus outage framework introduced in [25]. For this channel model

$$\rho = 1 - \exp\left(\frac{N_o W}{E[H]P} \left(2^{R/W} - 1 \right) \right), \quad (9)$$

where P is the transmission power, N_oW is the noise power, W is the bandwidth, and $E[H]$ is the expected value of the channel fading level, H. We assume that the fading is i.i.d. per packet. In our experiments, $N_oW/E[H]$ = 6Watts and W = 5MHz. We consider transmission rates ranging from R = 150Kbps to 300Kbps. These values are similar to the ones being proposed for next generation wireless standards [26]. It is important to note that the experimental setup is chosen to illustrate the concepts introduced in this paper and can be easily adapted based on the application and system requirements.



(a) (b)

Fig. 5. (a) Expected distortion per frame $\overline{D}^n$, (b) Average standard deviation in distortion per frame $\overline{\sigma}^n$ for the "silent" sequence.

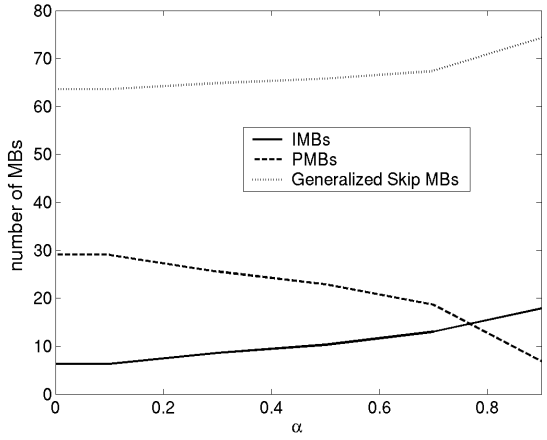


Fig. 6. Mode selection as a function of α for the "silent" sequence with R = 150Kbps and ρ = 0.01.

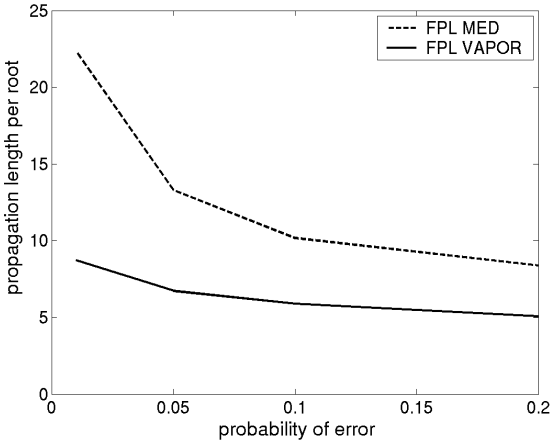


Fig. 7. Average error propagation length per root vs. probability of error for the "silent" sequence.

6.2. Error resilient source coding

In this section, we consider the problem of optimal source coding, i.e., the probability of packet loss is fixed. We focus on how adapting the source coding parameters, such as the prediction mode and quantizer, affects the end-to-end distortion.

Consider the "foreman" test sequence coded at 30 fps and transmitted over a 150 Kbps channel with probability of packet loss ρ = 0.01. In Fig. 4, we plot the expected distortion $\overline{D}_{seq}$ and the average standard deviation $\overline{\sigma}_{seq}$ for the sequence as a function of α. In other words, for each fixed value of α we solve (5) at each frame and plot the corresponding $\overline{D}_{seq}$ and $\overline{\sigma}_{seq}$. The motivation behind this experiment is to study how α affects the statistical properties

of the end-to-end distortion for the sequence. As shown in Fig. 4, $\overline{\sigma}_{seq}$ decreases as α increases. This is intuitively satisfying since a larger α means that more weight is placed on reducing the variance in distortion when allocation resources in (5).

Surprisingly though, the expected distortion for the sequence $\overline{D}_{seq}$ does not necessarily increase as α increases, as shown in Fig. 4. Recall that the optimization is carried out on a per frame basis and therefore ignores the effects of the current optimization on future frames. Therefore, although setting $\alpha = 0$ results in the lowest $\overline{D}^n$ for a single frame (as shown in Fig. 3), it is not guaranteed that $\overline{D}_{seq}$ is minimized by myopically minimizing the expected distortion per frame. This result is due to inter-frame dependencies. Reducing the variance in the current frame may lead to a more reliable prediction for the next frame, which in turn may reduce the overall expected distortion for the sequence.

Recently there has been work on prescient video coding, i.e., accounting for the affects on future frames when encoding the current frame [27]. The drawback of this work is that a certain number of frames must be captured before the optimization can begin, thus increasing delay. In addition, the reduction in overall distortion gained from optimizing over a group of frames is reported to be relatively small in [27]. Our results suggest that by reducing the variance of the distortion for the current frame we can achieve similar reductions in $\overline{D}_{seq}$ without the added delay or optimization complexity.

Next we consider the formulation in (7) where α is adapted per frame based on a tolerable increase in expected distortion Δ_d. Here we set $\Delta_d = 1.05$ (i.e., a 5% increase). In this experiment, we use the "silent" test sequence coded at 30 fps with R = 150 Kbps and ρ = 0.01. For these settings, the average value of α is 0.64. In Fig. 5 (a) and (b), we compare the proposed adaptive α approach with the MED approach (i.e., $\alpha = 0$). As shown in Fig. 5 (a) and (b), $\overline{D}^n$ per frame is similar for both approaches, while $\overline{\sigma}^n$ is significantly smaller for the VAPOR approach. This suggests that by accepting even a small increase in expected distortion it is possible to greatly reduce the variation in quality between different loss realizations. Similar results have been obtained for other sequences and parameter settings.

The value of α affects how much of the end-to-end distortion is due to source coding and how much is caused by channel errors. This is primarily a function of mode selection. For each value of α in (5), Fig. 6 shows the average number of MB's per frame that are coded as Intra (IMB's), Inter (PMB's), or are intentionally not transmitted (generalized skip). As α increases, the number of IMB's increases, as shown in Fig. 6. Because IMB's have lower coding efficiency than PMB's, the source coding distortion increases. On the other hand, increasing the number of IMB's reduces the distortion caused by error propagation. This trade-off will be discussed next.

6.3. Error Propagation

As in Sect. 6.2, we focus here on resilient source coding, i.e., mode and quantizer selection given a fixed probability of packet loss in the channel. We consider the "silent" sequence encoded at 30 fps and sent over a 150 Kbps channel. The objective in this section is to analyze how sensitive the MED and VAPOR approaches are to error propagation.

Let us define a pixel to be in error if its reconstructed value at the decoder differs from that at the encoder. At the receiver, we can track the temporal propagation of each error in order to identify error propagation paths. Each propagation path has a *root*, i.e., the origin of the error path, and a *length*, i.e., the number of pixel in error due to the initial root error. Note that only pixels that are lost and which start a new error path are defined to be root errors. In other words, a lost pixel which propagates a previous error is not considered to be a root error.

The average number of root errors per frame is a function of the probability of packet loss ρ. Since the probability of loss is fixed, both the MED and VAPOR approaches have roughly the same number of root errors per frame. As expected, the average number of roots per frame increases as ρ increases, for both approaches.

In Fig. 7, we plot the average length of error propagation per root as a function of ρ. The results are obtained by averaging over 50 channel loss simulations. As shown, the average number of pixel errors caused by an initial root error is significantly smaller for the VAPOR approach than the MED approach. This is especially true at lower probabilities of loss.

When the probability of loss is low, the expression for the expected distortion (2) is dominated by the expected distortion if the packet is received $E[D_R^{n,k,i}]$. In addition, the reference frame at the decoder is more likely to be correctly reconstructed, and hence source coding distortion becomes the primary component of $E[D_R^{n,k,i}]$. Therefore, as

channel conditions improve, an approach whose goal is to minimize the expected distortion will use more Inter coding in order to reduce the distortion due to compression (as shown in Fig. 6). The side affect of increased Inter coding is susceptibility to prolonged error propagation. This is why the average length of error propagation drastically increases for the MED approach as the probability of loss decreases, as seen in Fig. 7.

At low probability of loss, the expression for the variance in distortion (3) is dominated by $Var[D_R^{n,k,i}]$. As discussed earlier, Intra coding results in $Var[D_R^{n,k,i}] = 0$, while Inter coding has $Var[D_R^{n,k,i}] \geq 0$. Therefore, a variance-aware resource allocation scheme, such as (5) or (7), uses more Intra coding than a MED approach in order to reduce the variance in quality if a packet is received, as shown in Fig. 6. Perceptually, the increased number of Intra MBs results in faster termination of error propagation.

In Fig. 8, we compare a series of reconstructed frames at the decoder for the MED and VAPOR approaches. Note that these images are for a single channel loss realization and that the same MB's are lost in both schemes. As shown, both approaches suffer a loss in frame 109 where the woman's hand goes across her chin. The difference between the two approaches is that in frame 110, the VAPOR approach Intra refreshes this region while the MED approach does not. Thus, this error persists till frame 123 in the MED approach while it has been quickly removed by VAPOR. It is important to note that no feedback is used in either approach and that the difference in mode selection is purely due to estimates of the mean and variance of the end-to-end distortion.

From a communications point of view, transmitting information at the lowest possible probability of error is desirable. The point of the above discussion is that at this desirable operating point, an approach whose goal is to minimize the expected distortion may become susceptible to prolonged error propagation. A variance-aware approach on the other hand helps prevents prolonged error propagation because it accounts for the variability in quality caused by error propagation. As shown in Fig. 7, VAPOR reduces error propagation even at higher probabilities of packet loss. Thus, we argue that to be more resilient to error propagation, source coding techniques should account for both the mean and the variance of the distortion.

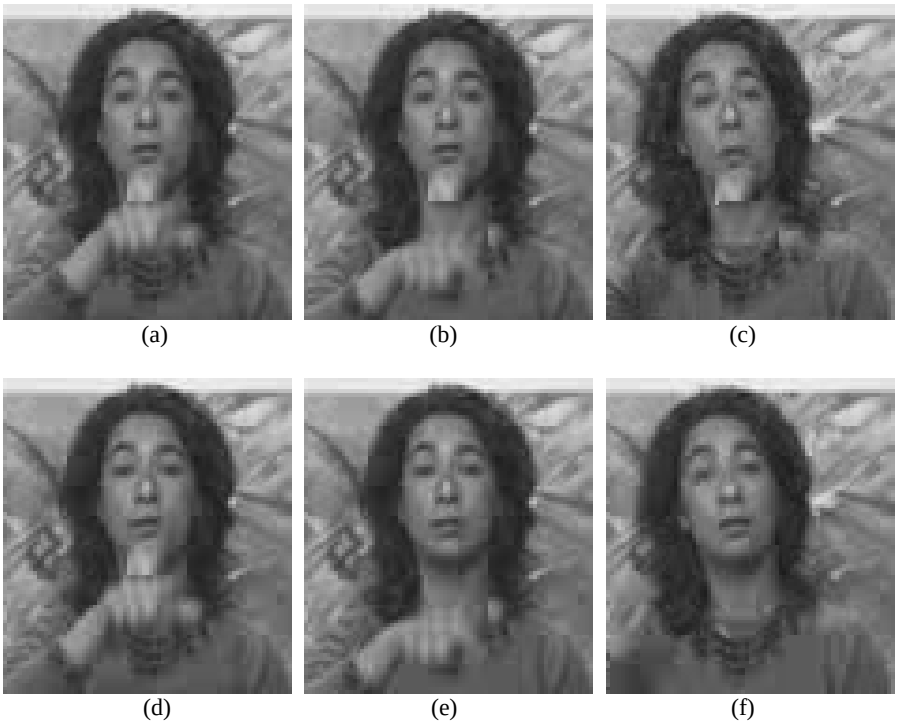


Fig. 8. Error propagation example. MED approach: frame # (a) 109, (b) 110, (c) 123.
VAPOR approach: frame # (d) 109, (e) 110, (f) 123.

7. CONCLUSION

This paper identifies the variance of the end-to-end distortion as an important quantity for characterizing video quality in packet lossy networks. A major contribution is the added flexibility and capability to control both the expected value and the variance of the distortion. Understanding human sensitivity to the different spatio-temporal artifacts caused by source and channel distortion is an area that requires significant research. This understanding will help determine the perceptual importance of the mean and the variance of the end-to-end distortion in video communication systems.

REFERENCES

[1] Y. Eisenberg, F. Zhai, C. E. Luna, T. N. Pappas, R. Berry, A. K. Katsaggelos, "Variance-aware distortion estimation for wireless video communications," *in Proc. ICIP*, Barcelona, Spain, September 2003.

[2] Y. Eisenberg, F. Zhai, T. N. Pappas, R. Berry, A. K. Katsaggelos, "VAPOR: Variance-Aware per-Pixel Optimal Resource-allocation," *IEEE Trans. Image Proc*, submitted January 2004.

[3] Y. Eisenberg, C. E. Luna, T. N. Pappas, R. Berry, and A. K. Katsaggelos, "Joint source coding and transmission power management for energy efficient wireless video communications," *IEEE Trans. CSVT*, vol. 12, no. 6, pp. 411-424, June 2002.

[4] Y. Wang, G. Wen, S. Wenger, and A. K. Katsaggelos, "Error resilient video coding techniques," *IEEE Signal Processing Mag*., vol. 17, pp. 61-82, July 2000.

[5] Y. Wang, Q.-F. Zhu, "Error control and concealment for video communication: A review," *Proc. IEEE*, pp. 974-997, May 1998.

[6] R.O. Hinds, T.N. Pappas, and J.S. Lim, "Joint block-based video source/channel coding for packet-switched networks," *Proc. SPIE*, vol. 3309, pp. 124-133, January 1998.

[7] T. Stockhammer, T. Wiegand, and S. Wenger, "Optimized transmission of H.26L/JVT coded video over packet-lossy networks," in *Proc. ICIP*, Rochester, New York, pp. 173-176, 2002.

[8] R.O. Hinds, "Robust mode selection for block-motion-compensated video encoding," Ph.D. thesis, MIT, Cambridge, MA, 1999.

[9] R. Zhang, S.L. Regunathan, and K. Rose, "Video coding with optimal Inter/Intra-mode switching for packet loss resilience," *IEEE JSAC*, pp. 966-976, June 2000.

[10] D. Wu, Y.T. Hou, B. Li, W. Zhu, Y.-Q. Zhang, and H. J. Chao, "An end-to-end approach for optimal mode selection in Internet video communication: theory and application," *IEEE J. Select. Areas Commun.*, vol. 18, pp. 977-995, June 2000.

[11] Leontaris and P. C. Cosman, "Video Compression with Intra/Inter Mode Switching and a Dual Frame Buffer," in *Proc. IEEE Data Compression Conference*, Snowbird, Utah, March 25-27, 2003.

[12] H. Yang and K. Rose, "Recursive end-to-end distortion estimation with model-based cross-correlation approximation," in *Proc. ICIP*, Barcelona, Spain, September 2003.

[13] G. Cote, S. Shirani, and F. Kossentini, "Optimal mode selection and synchronization for robust video communications over error-prone networks," *IEEE* JSAC, pp.952-965, June 2000.

[14] T. Wiegand, N. Farber, K. Stuhlmuller, and B. Girod, "Error-resilient video transmission using long-term memory motion-compensated prediction," *IEEE JSAC*, pp.1050-1062, June 2000.

[15] Z. He, J. Cai, and C. W. Chen, "Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding," *IEEE Trans. CSVT*, vol. 12, no. 6, pp. 511-523, June 2002.

[16] P. A. Chou and Z. Miao, "Rate-distortion optimized streaming of packetized media", *IEEE Trans. Mult.*, submitted Feb. 2001.

[17] G. M. Schuster, G. Melnikov, and A. K. Katsaggelos, "A Review of the Minimum Maximum Criterion for Optimal Bit Allocation Among Dependent Quantizers," *IEEE Trans. on Multimedia*, vol. 1, no. 1, pp. 3-17, March 1999

[18] Y. Eisenberg, C. E. Luna, T. N. Pappas, R. Berry, and A. K. Katsaggelos, "Optimal source coding and transmission power management using a min-max expected distortion approach," in *Proc. ICIP*, Rochester, New York, pp. 537-540, 2002.

[19] C. Y. Hsu, A. Ortega, and M. Khansari, "Rate control for robust video transmission over burst-error wireless channels," *IEEE J. Select. Areas Comm.*, vol. 17, pp. 756-773, May 1999.

[20] A. Munteanu, Y. Andreopoulos, M. van der Schaar, P. Schelkens, and J. Cornelis, "Control of the distortion variation in video coding systems based on motion compensated temporal filtering," *in Proc. ICIP*, Barcelona, Spain, September 2003.

[21] G.M. Schuster and A.K. Katsaggelos, *Rate-Distortion Based Video Compression: Optimal Video Frame Compression and Object Boundary Encoding*, Kluwer, 1997.

[22] Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Mag.*, vol. 15, pp. 23-50, November 1998.

[23] D. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1995.

[24] F. Zhai, Y. Eisenberg, T. N. Pappas, R. Berry, and A. K. Katsaggelos, "Joint Source-Channel Coding and Power Allocation for Energy Efficient Wireless Video Communications," *in Proc. Allerton*, Monticello, Illinois, October 2003.

[25] L. Ozarow, S. Shamai, and A. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Technol.*, vol. 43, No. 2, pp. 359-378, May 1994.

[26] S. Nanda, K. Balachandran, and S. Kumar, "Adaption techniques in wireless packet data services," *IEEE Commun. Mag.*, vol. 38, pp. 54-64, January 2000.

[27] R. Zhang, S.L. Regunathan, K. Rose, "Prescient mode selection for robust video coding," *in Proc. ICIP*, pp. 974-977, 2001.