# VARIANCE-AWARE DISTORTION ESTIMATION FOR WIRELESS VIDEO COMMUNICATIONS

*Yiftach Eisenberg, Fan Zhai, Carlos E. Luna,*
*Thrasyvoulos N. Pappas, Randall Berry, and Aggelos K. Katsaggelos*

Northwestern University, Department of ECE, Evanston, IL 60208, USA
E-mail: {yeisenbe, fzhai, carlos, pappas, rberry, aggk}@ece.northwestern.edu

## ABSTRACT

The problem of encoding and transmitting a video sequence over a wireless channel is considered. Our objective is to minimize the end-to-end distortion while using a limited amount of transmission energy and delay. In our approach, we jointly adapt the source-coding parameters and transmission power per packet. We introduce the concept of "Variance-Aware Distortion Estimation" (VADE), and present a framework for controlling both the expected value and the variance of the end-to-end distortion. This framework is based on knowledge of how the video is compressed, the probability of packet loss, and the concealment strategy. To the best of our knowledge, this paper is the first to address the trade-off between the mean and variance of the end-to-end distortion. Experimental results demonstrate the potential of the proposed approach.

## 1. INTRODUCTION

Transmission energy is a critical resource in wireless video communications [1]. Since most users of a wireless network are mobile, they must rely on a battery with a limited energy supply. Efficiently utilizing transmission energy can extend the lifetime of this battery, decrease the level of interference between users, as well as increase the overall network capacity. This paper builds on our prior work, some of which can be found in [2]. Our goal is to achieve the best video quality using a limited amount of transmission energy and delay. To accomplish this we jointly consider error resilience and concealment techniques at the source-coding level, and transmission power management at the physical layer. In this way, the transmission power/energy is used as an unequal error protection (UEP) mechanism.

In most video communication systems, the transmitter does not know exactly which packets are lost, but instead has an estimate of the probability of packet loss. Thus, from the point of view of the transmitter, the distortion at the receiver is a random variable. Recent work on resilient video coding for packet loss networks has primarily focused on minimizing the expected value of the end-to-end distortion [2,3,4,5,6]. A common feature among these works is that they all measure video quality by the expected distortion, where the expectation is computed with respect to all the possible packet loss patterns.

Several methods have been proposed for calculating the expected distortion. These methods can be divided into two general categories. The first is optimal per-pixel estimation methods, such as [3,4,2], that can be used to accurately calculate the expected value of the distortion under certain conditions.

(a) (b) (c)

Fig. 1. (a) Expected frame, (b and c) two loss realizations.

The second category consists of methods that use models to estimate the expected distortion [5,6]. Model based methods are useful when either computation power is limited or closed form expressions for the expected distortion are not known. The above is only a small sample of the work in this area.

At the receiver, the end user sees only one of the many possible reconstructed sequences, depending on which packets are lost. Therefore, the actual distortion at the receiver is not equal to 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, the quality at the receiver may vary greatly based on which packets are lost. Therefore, in this paper we argue that the variance of the end-to-end distortion should also be considered when characterizing video quality in lossy packet networks. We introduce the concept of "variance-aware distortion estimation" (VADE), and present a framework for controlling both the expected value and the variance of the end-to-end distortion.

## 2. SYSTEM MODEL

Consider a video communication system where the video is encoded using a block-based motion-compensated technique (e.g., H.263, MPEG-4). Each frame is divided into slices that are comprised of consecutive Macro-Blocks (MBs). Each slice is independently decodable, i.e., the decoding of one slice is not affected by the loss of other slices in the same frame. Losses in other frames may cause temporal error propagation due to inter-frame prediction. After a slice is encoded, it is transmitted across a wireless channel as a separate packet. In the following, slice and packet will be used interchangeably. Let $M$ be the number of packets in a given frame and $k$ be the packet index.

For each packet, source-coding parameters, such as the coding mode (intra/inter/skip) and quantization step-size for each MB are specified. We use $\mu^k$ to denote the source-coding parameters for the $k$th packet, and $\boldsymbol{\mu} = \{\mu^1 \ldots \mu^M\}$ to denote the coding parameters for all the packets in a frame. The number of bits used to encode the $k$th packet, $B^k$, is a function of $\mu^k$; we use $B^k(\mu^k)$ to explicitly indicate this dependency.

In addition to $\mu^k$, we assume that the transmission power for each packet, $P^k$, can be adjusted. We use $\boldsymbol{P} = \{P^1 \ldots P^M\}$ to denote the transmission power for all the packets in a frame.

Similarly, let $\rho^k$ denote the probability of loss for the $k$th packet, and $\boldsymbol{\rho} = \{\rho^1 \ldots \rho^M\}$. We assume that a function $f$, relating $P^k$ to $\rho^k$ is known at the transmitter, i.e., $\rho^k = f(P^k)$. This function can be determined from empirical measurements or analytical models; we provide one example in Sect. 6. The work presented here is applicable to any system where the relationship between transmission power and the probability of loss can be found.

The total energy used to transmit all the packets in a frame is

$$\mathrm{E}_{tot} = \sum_{k=1}^{M} B^k(\mu^k)\frac{P^k}{R^k}, \tag{1}$$

where $R^k$ is the transmission rate for the $k$th packet in encoded video bits per second. Notice that $\mathrm{E}_{tot}$ is a function of $\mu^k$ and $P^k$. This is one reason why we consider jointly adapting these parameters.

## 3. END-TO-END DISTORTION

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

Consider a single pixel in the video sequence whose original value is $x$. 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, as shown in Fig. 2. In this case, the reconstructed pixel value at the receiver is a random variable that can be expressed as

$$Y = \begin{cases} Y_R & \text{w/ prob } (1\text{-}\rho) \\ Y_L & \text{w/ prob } (\rho) \end{cases}, \tag{2}$$

where $\rho$ is the probability that the packet containing the coding parameters for this pixel is lost, $Y_R$ is the reconstructed pixel value if the packet is *received* correctly, and $Y_L$ is the reconstructed value if the packet is *lost*. If the pixel is predictively encoded, then $Y_R$ is a random variable. On the other hand, if the pixel is independently encoded (Intra coded), then $Y_R$ is deterministic. $Y_L$ depends on the concealment strategy. If prediction is used in the concealment strategy, e.g., temporal concealment, then $Y_L$ is also a random variable.

We assume that a known distortion metric $D(x,Y)$ is used to evaluate the distortion in the reconstructed pixel. For example, in Sect. 6 we use the squared error metric, i.e., $D(x,Y) = (x\text{-}Y)^2$. It is important to note that any distortion metric that maps the pair $(x,Y)$ to a distortion value $D$ can be used in this framework. The distortion between $x$ and $Y$ can be expressed as

$$D(x,Y) = \begin{cases} D_R(x,Y_R) & \text{w/ prob } (1\text{-}\rho) \\ D_L(x,Y_L) & \text{w/ prob } (\rho) \end{cases}, \tag{3}$$

where $D_R(x,Y_R)$ and $D_L(x,Y_L)$ are the distortion if the source-coding parameters are received and lost respectively. In order to simplify notation, we abbreviate $D(x,Y)$ as $D$, $D_R(x,Y_R)$ as $D_R$, and $D_L(x,Y_L)$ as $D_L$. The probability mass function (pmf) of $D$ and the relevant quantities discussed below are shown in Fig. 3.

### 3.1. Expected end-to-end distortion

The expected value of the end-to-end distortion for a given pixel is by definition

$$E[D] = (1\text{-}\rho)E[D_R] + (\rho)E[D_L], \tag{4}$$

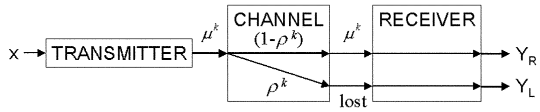



Fig. 2. System model

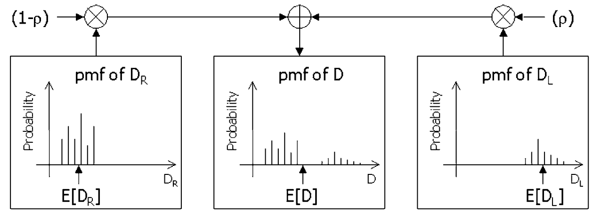



Fig. 3. pmf of $D$, $D_R$, and $D_L$ and their relationship

where $E[\bullet]$ indicates the expected value. Different channel loss realizations can cause different distortion at the receiver. $E[D]$ is a measure of the average distortion for a given pixel, where the average is taken with respect to all the possible loss realizations. One way to reduce $E[D]$ is by allocating more source-coding resources to this pixel in order to decrease $E[D_R]$. Alternatively, allocating more communication resources, e.g., transmission power, can lower $\rho$, thus decreasing $E[D]$ (Here we assume that $E[D_R] < E[D_L]$). The average expected distortion for the $k$th packet, $E[D^k]$, is obtained by simply taking the average $E[D]$ for all the pixels in the $k$th packet.

### 3.2. Variance of the end-to-end distortion

The variance of the end-to-end distortion is also important when evaluating video quality in communication systems. The variance in distortion indicates the likelihood that the actual distortion in the reconstructed sequence (i.e., for a single loss realization) is close to $E[D]$. Therefore, the variance of the end-to-end distortion is a measure of how reliable $E[D]$ is as an estimate of the distortion at the receiver.

The variance of the end-to-end distortion for a given pixel is by definition $Var[D] = E[D^2] - E[D]^2$. By substituting (3) into the previous equation and rearranging terms, we can express $Var[D]$ as

$$\begin{aligned} Var[D] \; &= \; (1\text{-}\rho)Var[D_R] + (\rho)Var[D_L] \\ &\quad + (1\text{-}\rho)(\rho)\{E[D_R] - E[D_L]\}^2, \end{aligned} \tag{5}$$

where $Var[D_R]$ and $Var[D_L]$ are the variance in distortion if the packet is received and lost, respectively. As expected, $Var[D]$ increases when $Var[D_R]$ or $Var[D_L]$ increase. Therefore, Intra coding, which has $Var[D_R] = 0$, enables the transmitter to decrease $Var[D]$. Inter coding on the other hand has $Var[D_R] \geq 0$. From (5) we see that $Var[D]$ increases as $\{E[D_R]\text{-}E[D_L]\}^2$ increases. This means that if a pixel is difficult to conceal, i.e, $E[D_L] \gg E[D_R]$, its distortion may vary greatly depending on whether the packet is received or lost.

$Var[D]$ is a negative quadratic function of $\rho$, as shown in (5). Thus, $Var[D]$ has a maximum at $\rho = 1/2 - \{Var[D_L]\text{-}Var[D_R]\}/2\{E[D_R]\text{-}E[D_L]\}^2$, and decreases as $\rho$ increases or decreases from this value. This is intuitively satisfying since there is less variability in $D$ when $\rho$ is either very small or very large. For example, when $\rho = 0$ or $1$ for all the packets in the sequence, $Var[D] = 0$. The average variance in distortion for the $k$th packet, $Var[D^k]$, is obtained by simply taking the average of

$Var[D]$ for all the pixels in the $k$th packet. Similarly, the average standard deviation in distortion for the $k$th packet, $Std[D^k]$, is equal to the average of $Std[D] = \sqrt{Var[D]}$. It is important to note that the average of the variance in distortion per pixel is not equal to the variance of the average distortion for a frame. The latter is more difficult to calculate and does not capture local variation in quality well.

Consider the case where $D = (x\text{-}Y)^2$ (squared error distortion metric). In this case, we need the first two moments of $Y$, i.e., $E[Y]$ and $E[Y^2]$, in order to accurately calculate $E[D]$. Similarly, the first four moments of $Y$ are needed to accurately calculate $Var[D]$. In certain cases, recursively optimal distortion estimation methods, such as ROPE [4] and [3] can be used to efficiently calculate the necessary reconstructed pixel moments.

## 4. VARIANCE-AWARE FORMULATION

In the previous sections we demonstrated the need to account for both the mean and the variance of the distortion when evaluating video quality in packet loss environments. One way to do this is by defining the distortion for a given frame, $D_{tot}$, as the average weighted sum of $\{(1\text{-}\alpha)E[D^k] + (\alpha)Std[D^k]\}$ for all $k$. We use $Std[D^k]$ instead of $Var[D^k]$ so that the units of $D_{tot}$ are consistent. Other expressions for $D_{tot}$ that incorporate both the mean and variance of the distortion can also be used.

Our goal is to control both the source-coding and transmission power in order to minimize the end-to-end distortion while using a limited amount of transmission energy and delay. We can formally write this optimization as

$$\min_{\{\mu^k, P^k\ \forall k\}} D_{tot} = \frac{1}{M}\sum_{k=1}^{M}(1-\alpha)E[D^k(\boldsymbol{\mu},\boldsymbol{\rho})] + (\alpha)Std[D^k(\boldsymbol{\mu},\boldsymbol{\rho})] \quad (6.a)$$

$$\text{s.t.:}\quad \mathrm{E}_{tot} = \sum_{k=1}^{M} B^k(\mu^k)\frac{P^k}{R^k} \le \mathrm{E}_0 \quad (6.b)$$

$$T_{tot} = \sum_{k=1}^{M}\frac{B^k(\mu^k)}{R^k} \le T_0\,, \quad (6.c)$$

where $\mathrm{E}_{tot}$ is the total transmission energy, $\mathrm{E}_o$ is the transmission energy constraint, $T_{tot}$ is the total transmission delay, and $T_o$ is the transmission delay constraint for the frame. In this work we assume that the transmission rate is fixed, i.e., $R^k = R$ for all $k$. The formulation can easily be extended to allow variable $R^k$ per packet. We refer to the formulation in (6) as a "Variance-Aware per-Pixel Optimal Resource-allocation" (VAPOR) technique. We assume that a higher-level controller assigns energy and delay constraints per frame based on the application. The design of this controller is an area of future research.

In (6.a), $\alpha$ is used to control the relative importance of the variance in the end-to-end distortion. Increasing $\alpha$ results in a smaller variance in distortion. To the best of our knowledge, the formulation in (6) is the first to account for both the mean and the variance of the end-to-end distortion. When $\alpha = 0$, we obtain a special case of the general formulation in which the objective is to minimize the average expected distortion, as in [3-6]. When $P^k = P\ \forall k$, each packet is transmitted with a fixed power, and thus the same level of protection. We refer to this special case as the "Fixed Power" (FP) approach. When $P^k$ is chosen from a set**,** we have the "Variable Power" (VP) approach in which the probability of loss can be adjusted per packet by adapting the transmission power. In [2], optimal source-coding and transmission power management for energy efficient wireless video communications is considered in detail.

## 5. SOLUTION APPROACH

In order to solve (6) we use Lagrangian relaxation and dynamic programming (DP). First we introduce two Lagrange multipliers, $\lambda_1 \ge 0$ and $\lambda_2 \ge 0$, and solve the relaxed problem

$$\min_{\{\mu^k, P^k\ \forall k\}} \quad D_{tot} + \lambda_1 \mathrm{E}_{tot} + \lambda_2 T_{tot}\,. \quad (7)$$

By appropriately choosing $\lambda_1$ and $\lambda_2$, the solution to (6) can be obtained within a convex-hull approximation by solving (7). Various methods, such as cutting-plane or sub-gradient methods, can be used to search for $\lambda_1$ and $\lambda_2$ [7]. In our experimental results, we use an efficient method developed in [8] that exploits the structure of the problem presented in (6).

For each choice of $\lambda_1$ and $\lambda_2$ we can solve (7) using dynamic programming. 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. In (6.a), we represent these dependencies by indicating that $E[D^k]$ and $Std[D^k]$ may depend on $\boldsymbol{\mu}$ and $\boldsymbol{\rho}$. Dynamic programming can be used to efficiently find the optimal $\mu^k$ and $P^k$ for each packet in the frame when the dependencies between packets are limited, e.g., to a small neighborhood. For more details please see [2, 9, 10].

## 6. EXPERIMENTAL RESULTS

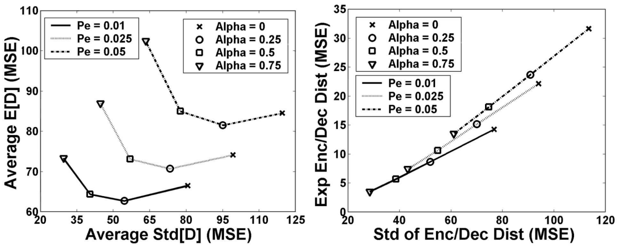
In our experiments, we use the "Foreman" sequence in QCIF format encoded at 30 frames per second using MPEG-4. A limited number of quantization step sizes are used for "Intra" and "Inter" MBs. We consider the case where each packet contains a single MB, i.e., each MB is independently decodable. This packetization scheme has a low coding efficiency but helps illustrate the concepts introduced in this paper. 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.

Consider the case where each packet is sent over a narrow-band slowly fading channel with additive white Gaussian noise. We model $\rho^k$ in the capacity versus outage framework introduced in [11], and assume that the channel fading is i.i.d. per packet. For this channel model

$$\rho^k = 1 - \exp\left(\frac{N_o W}{P^k E[H]}\left(2^{R/W} - 1\right)\right), \quad (8)$$

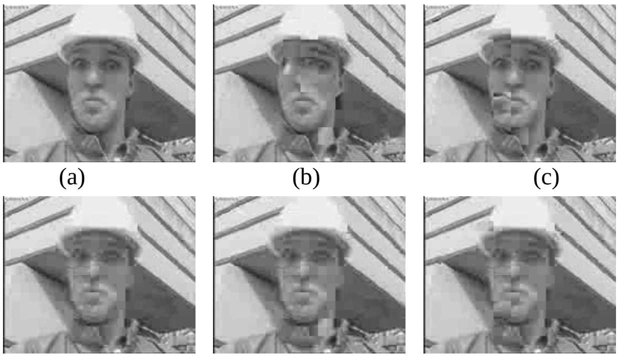
where $N_o W$ is the noise power, $W$ is the bandwidth, and $E[H]$ is the expected value of the channel fading level, $H$. In our experiments, $N_o W/E[H]$ = 6Watts, $W$ = 5MHz, and $R$ = 150Kbps. These values are similar to the ones being proposed for next generation wireless standards [12]. We consider a real-time application with an allowable transmission delay of one frame duration, i.e., $T_0$ = 33 msec.

In Fig. 4, we show how the expected value and the standard deviation of the distortion, averaged over the entire "Foreman" sequence, are affected by the value of $\alpha$ in (6). We consider the FP approach with several fixed packet loss probabilities. In Fig. 4(a), we analyze the distortion between the original and the reconstructed sequence at the decoder. Notice that as $\alpha$ increases, the average $Std[D]$ consistently decreases.




(a) (b)

Fig. 4. (a) Avg. E[D] vs. Std[D] between Original and Decoded
(b) Avg. E[D] vs. Std[D] between Encoded and Decoded



(a) (b) (c)

(d) (e) (f)

Fig. 5. FP with $\alpha$=0: (a) encoded frame, (b,c) loss realizations.
VAVP with $\alpha$=1: (d) encoded frame, (e,f) loss realizations.

Surprisingly, the average $E[D]$ does not necessarily increase as $\alpha$ increases. For example, consider the curve with $\rho = 0.01$. In this case, setting $\alpha = 0.25$ results in an average $Std[D]$ that is 32% lower than that achieved for $\alpha = 0$, as well as an average $E[D]$ that is 6% lower. This result is likely due to inter-frame dependencies. Reducing the variance in the current frame may lead to a more reliable prediction for the next frame. This in tern may reduce the overall distortion for the sequence. Thus, it may be possible to significantly reduce the average variance in distortion per pixel without sacrificing the expected distortion.

In Fig. 4 (b), we examine the distortion between the reconstructed sequence at the encoder and the decoder. As $\alpha$ increases, the expected value and standard deviation both decrease, i.e, the reconstructed sequence at the decoder more closely resembles what was transmitted. Therefore, by controlling the variance we can reduce what is sometimes referred to as channel distortion. In Fig. 5, we compare a fixed power (FP) approach with a variance-aware variable power (VAVP) approach. For the FP approach, we optimally select the source coding parameters in order to minimize the expected distortion, assuming a fixed probability of loss, $\rho = 0.20$ (as in [3-6]). In the VAVP approach, we jointly adapt the source coding and the transmission power in order to minimize (6.a), with $\alpha = 0.5$. Both approaches have the same energy and delay constraints, (6.b) and (6.c). For these settings, the VAVP approach achieves an average $E[D]$ and average $Std[D]$ that are 21% and 39% smaller than those achieved by the FP approach. This suggests that by jointly adapting the coding parameters and transmission power, and by incorporating the variance into the distortion evaluation, we can significantly decrease both the mean and the variance of the end-to-end distortion.

As shown in Fig. 5 (b,c) and (e,f), the reconstructed frames at the receiver, for two different loss realizations, more closely resemble the encoded frame in the VAVP approach. The tradeoff for lower channel distortion is a possible increase in source coding distortion, as shown in Fig. 5 (a,d). Understanding the complex 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 distortion.

## 7. CONCLUSIONS

This paper identifies the variance in distortion as an important quantity when characterizing video quality in packet loss environments. A major contribution is the added flexibility and capability to control both the expected value and the variance of the end-to-end distortion. In addition, the concepts introduced in this paper can also be extended to other cost-distortion optimization problems, such as video over differentiated services networks [8].

## 8. REFERENCES


[1] N. Bambos, "Toward power-sensitive network architectures in wireless communications: concepts, issues, and design aspects," *IEEE Pers. Commun.*, pp. 50-59, June 1998.

[2] 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*, pp. 411-424, June 2002.

[3] 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.

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

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

[6] 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.

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

[8] F. Zhai, C.E. Luna, Y. Eisenberg, T.N. Pappas, R. Berry, and A.K. Katsaggelos, "Joint source coding and packet classification for video streaming over differentiated services networks," *IEEE TMM*, submitted January 2003.

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

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

[11] 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.

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