Optimal Mode Selection in Internet Video Communication: An End-to-End Approach

Dapeng Wu* Y. Thomas Hou† Ya-Qin Zhang‡ H. Jonathan Chao§

Abstract

We present an *end-to-end approach* to generalize the classical theory of R-D optimized mode selection for point-to-point video communication. We introduce a notion of *global* distortion by taking into consideration of both the path characteristics and the receiver behavior, in addition to the source behavior. We derive, for the first time, a set of accurate global distortion metrics for any packetization scheme. Equipped with the global distortion metrics, we design an R-D optimized mode selection algorithm to provide the best trade-off between compression efficiency and error resilience. As an application, we integrate our theory with point-to-point MPEG-4 video conferencing over the Internet. Simulation results conclusively demonstrate that our end-to-end approach offers superior performance over the classical approach for Internet video conferencing.

1 Introduction

Video communication over the Internet is becoming an important application in recent years. For video communication over a network, a coding algorithm such as H.263 or MPEG-4 [3] usually employs rate control to match the output rate to the available bandwidth. This can be achieved by choosing a mode that minimizes the quantization distortion between the original frame/macroblock and the reconstructed one under a given bit budget [5, 10], which is the so-called rate-distortion (R-D) optimized mode selection. We refer such R-D optimized mode selection as the *classical approach*. The classical approach is not able to achieve global optimality under the error-prone environment since it does not consider the network congestion status and the receiver behavior.

This paper presents an *end-to-end approach* to solve the fundamental problem of R-D optimized mode selection for peer-to-peer video communication over packet-switched networks. We identify three factors that have impact on the video presentation quality at the receiver, namely, *the source behavior, the path characteristics,* and *the receiver behavior.* We formulate the problem of globally optimal mode selection using a new notion of global distortion metric and derive, for the first time, a set of accurate global distortion metrics for any packetization scheme. Equipped with the global distortion metrics, we design an R-D optimized mode selection algorithm to provide the best trade-off between compression efficiency and error resilience. Our theory on R-D optimized mode selection is general and is applicable to many video coding standards, including H.261/263 and MPEG-1/2/4.

As an application, we apply our global R-D optimized mode selection to point-to-point MPEG-4 video conferencing over the Internet. Simulation results conclusively demonstrate that our end-to-end approach offers superior performance over the classical approach for Internet video conferencing.

The remainder of this paper is organized as follows. Sections 2 and 3 present the theory of globally optimal mode selection. As an application, Section 4 presents an end-to-end implementation architecture for point-to-point MPEG-4 video conferencing over the Internet, the performance of which is demonstrated with simulation results in Section 5. Section 6 concludes this paper.

2 An End-to-End Approach

We organize this section as follows. In Section 2.1, we introduce the notion of global distortion and formulate the problem of globally optimal mode selection. Section 2.2 examines the key factors contributing to the global distortion.

2.1 Problem Formulation

For Internet video communication, on the sender side, raw bit-stream of live video is encoded by a video encoder. After this stage, the compressed video bit-stream is first packetized and then passed through the transport protocol layers before entering the network. Packets may be dropped inside the network (due to congestion) or at the destination (due to excess delay). For packets that are successfully delivered to the destination, they first pass through the transport protocol layers and depacketized before being decoded at the video decoder.

Table 1 lists the notations used in this paper [14].

In formulating the problem of globally R-D optimized mode selection, we consider an MB at location i ($i \in [0, N_h]$) of a given frame. We assume that each MB can be coded using only one of the two modes in $\mathcal{I}$.

The problem of classical R-D optimized mode selection is to find the mode that minimizes the quantization distortion D_q for a given MB, subject to a constraint R_c on the number of bits used. This constrained problem can be formulated as

$$\min_{M_i^n} D_q(M_i^n) \quad \text{subject to} \quad R(M_i^n) \leq R_c,$$

where $D_q(M_i^n)$ and $R(M_i^n)$ denote the quantization distortion and the number of bits used, respectively, for macroblock F_i^n with a particular mode M_i^n.

*D. Wu is with Polytechnic University, Dept. of Electrical Engineering, Brooklyn, NY, USA.

†Y. T. Hou is with Fujitsu Laboratories of America, Sunnyvale, CA, USA.

‡Y.-Q. Zhang is with Microsoft Research, Beijing, China.

§H. J. Chao is with Polytechnic University, Dept. of Electrical Engineering, Brooklyn, NY, USA.

0-7803-6283-7/00/$10.00 © 2000 IEEE

Table 1: Notations.

N_f	the total number of MBs in a frame.
N_h	the highest location number of MBs in a frame ($N_h = N_f - 1$).
N_G	the number of MBs in a group of blocks (GOB).
F_i^n	the MB at location i in frame n.
$\bar{F}_i^n$	the coded MB at location i in frame n.
$F_{\tilde{i}}^n$	the MB (at location $\tilde{i}$ in frame n) which is above F_i^n, if it exists.
$\mathcal{G}^n$	the set of macroblocks $\bar{F}_i^n$ ($i \in [0, N_h]$) that does not have $F_{\tilde{i}}^n$.
$\hat{P}_R^{(i,n)}$	the probability of the event that $\bar{F}_i^n$ is received correctly.
$\hat{P}_L^{(i,n)}$	the probability of the event that $\bar{F}_i^n$ is lost.
$\hat{P}_{RL}^{(i,n)}$	the probability of the event that $\bar{F}_{\tilde{i}}^n$ is received correctly and $\bar{F}_i^n$ is lost.
$\hat{P}_{LL}^{(i,n)}$	the probability of the event that both $\bar{F}_i^n$ and $\bar{F}_{\tilde{i}}^n$ are lost.
f_{ij}^n	the original value of pixel j in F_i^n (raw data).
$\tilde{f}_{ij}^n$	the value of reconstructed pixel j in F_i^n at the encoder.
$\hat{f}_{ij}^n$	the value of reconstructed pixel j in F_i^n at the receiver.
e_{ij}^n	the prediction error of pixel j in inter-coded F_i^n.
$\tilde{e}_{ij}^n$	the reconstructed prediction error of pixel j in inter-coded F_i^n.
$\tilde{f}_{uv}^{n-1}$	the value of reconstructed pixel v in F_u^{n-1} for prediction of f_{ij}^n.
$\hat{f}_{ml}^{n-1}$	the value of reconstructed pixel l in F_m^{n-1} to replace $\hat{f}_{ij}^n$ due to *EC*-3.
$\mathcal{I}$	the set of coding modes (i.e., $\mathcal{I}$ = {intra, inter}).
M_i^n	the mode selected to code macroblock F_i^n ($M_i^n \in \mathcal{I}$).
X_k	the packet with sequence number k ($k \geq 0$).
η_i^n	the sequence number of the last packet used to packetize macroblock $\bar{F}_i^n$.
$\prec$	the completely containing relation between a macroblock and a packet.
$\preceq$	the partially containing relation between a macroblock and a packet.
$\mathcal{X}$	the set of packets that packetize frame 0, i.e., $\mathcal{X} = \{X_{\eta_i^0} : i \in [0, N_h]\}$.
K	the number of packets in set $\mathcal{X}$.

The classical R-D optimized mode selection is optimal with respect to quantization distortion. However, the classical R-D optimized mode selection is not optimal with respect to the distortion D_r, which measures the difference between the original image/frame/MB at the source and the reconstructed one at the receiver. This is because the classical R-D optimized mode selection does not consider the path characteristics (packet loss) and receiver behavior (error concealment), both of which affect the distortion D_r. This motivates us to propose globally R-D optimized mode selection.

We consider the distortion D_r, which is the difference between the original image/frame/MB at the source and the reconstructed one at the receiver. Under lossy environments such as Internet and wireless communication, the distortion D_r is a random variable, which may take the value of either (1) the quantization distortion D_q plus the distortion D_{ep} caused by error propagation, or (2) distortion D_c caused by errors due to error concealment. We define the *global* distortion D as the expectation of the random variable D_r. That is,

$$D = E\{D_r\},$$

where D_r takes the value of $(D_q + D_{ep})$ or D_c with certain probability, which is determined by path characteristics (packet loss behavior). Therefore, the global distortion takes three factors into consideration: sender behavior (quantization and packetization), path characteristics, and receiver behavior (error concealment).

The problem of globally R-D optimized mode selection is to find the mode that minimizes the global distortion D for a given MB, subject to a constraint R_c on the number of bits used. This constrained problem reads as follows.

$$\min_{M_i^n} D(M_i^n) \quad \text{subject to} \quad R(M_i^n) \leq R_c,$$

where $D(M_i^n)$ denotes the global distortion for macroblock F_i^n with a particular mode M_i^n.

The global distortion can be expressed by the sum of absolute differences (SAD), mean absolute difference (MAD), the sum of squared differences (SSD), mean squared error (MSE), or peak signal-to-noise ratio (PSNR). In this paper, we define the global distortion metrics for macroblock F_i^n in terms of MAD as follows.

$$\text{MAD}(F_i^n) = \frac{E\left\{\sum_{j=1}^{256} |f_{ij}^n - \hat{f}_{ij}^n|\right\}}{256}.$$

For the rest of the paper, we will develop theory based on MAD. However, our underlying methodology is general and can be applied to other global distortion metrics (i.e., SAD, SSD, MSE, PSNR).

2.2 Key Factors in the Global Distortion Metric

This subsection discusses in detail the three factors (i.e., sender behavior, path characteristics, and receiver behavior) that contribute to the global distortion.

2.2.1 Source Behavior

The source behavior includes quantization and packetization, both of which have impact on global distortion. A description of the video encoder can be found in [14] and we will thus focus our discussion on packetization part of the source behavior.

Throughout the paper, we assume that the payload size of a packet is always greater than the size of any MB. Note that under such assumption, an MB could be split into no more than two consecutive packets. We define several packetization schemes as follows.

Definition 1 (Packetization Schemes) *A packetization scheme for video bit-stream is called*

- *PKT-1 if each generated packet has the same fixed packet size;*
- *PKT-2 if each generated packet solely contains a complete MB;*
- *PKT-3 if each generated packet solely contains a complete GOB/slice.*

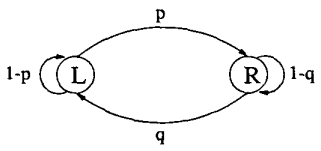

Figure 1: Gilbert path model.

Under PKT-1, the packet size can be set as large as path MTU to achieve efficiency. PKT-1 is widely used due to its simplicity. If PKT-1 is used, there are only two cases for the relation between an MB and the last packet used to packetize it: (1) an MB ($\bar{F}_i^n$) is completely contained by a single packet ($X_{\eta_i^n}$), i.e., $\bar{F}_i^n \prec X_{\eta_i^n}$; or (2) an MB ($\bar{F}_i^n$) is split into two consecutive packets and partially contained by each packet ($X_{\eta_i^n}$),[1] i.e., $\bar{F}_i^n \preceq X_{\eta_i^n}$.

If PKT-2 is used, an MB never gets split into two packets. Thus, loss of a packet only corrupts one MB, which enhances the error resilient capability of the video. For this reason, PKT-2 was adopted by Internet Engineering Task Force (IETF) [11].

PKT-3 has similar property to that of PKT-2. That is, a GOB/slice/MB is never split into two packets.[2] Thus, loss of a packet only corrupts one GOB/slice. PKT-3 was also adopted by IETF [15].

2.2.2 Path Characteristics

Measurements of packet loss in the Internet have shown that the packet loss behavior can be modeled reasonably well with a 2-state Markov chain, also known as *the Gilbert model* (see Fig. 1) [1]. That is, the Markov chain is in state R if a packet is received timely and correctly and in state L if a packet is lost either due to network congestion or due to exceeding the maximum delay threshold. The parameters p and q are the transition probabilities between states L and R. The durations of states L and R are exponentially distributed with respective mean lengths T_L and T_R, which are given by $T_L = \frac{1}{p}$ and $T_R = \frac{1}{q}$, respectively. The probability of the event that the path is in state L (i.e., packet loss probability) is given by $P_L = \frac{T_L}{T_R+T_L} = \frac{q}{p+q}$. The transition matrix $\mathbf{A}$ of the 2-state Markov chain is given by

$$\mathbf{A} = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}.$$

In the end-to-end architecture, the transition probabilities p and q are given by:

$$p = \frac{N_1}{N_1+N_2} \quad \text{and} \quad q = \frac{N_3}{N_3+N_4}, \qquad (1)$$

where N_1 is the number of successfully received packets when the previous packet is lost, N_2 is the number of lost packets when the previous packet is lost, N_3 is the number of lost packets when the previous packet is successfully received, and N_4 is the number of successfully received packets when the previous packet is successfully received.[3] N_1, N_2, N_3, and N_4 can be measured at the receiver.

[1] To be specific, if $\bar{F}_i^n$ is split, $\bar{F}_i^n$ is partially contained by packet $X_{\eta_i^n-1}$ and partially contained by packet $X_{\eta_i^n}$.

[2] MPEG-4 does not have the concept of GOB for video sequences with arbitrary shape. However, we can define a slice which is the part of GOB confined by two shape boundaries of the VO.

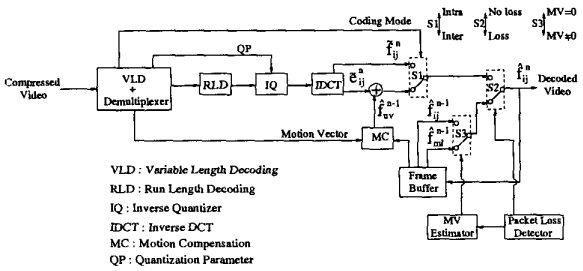

Figure 2: Block diagram of the video decoder.

2.2.3 Receiver Behavior

We define several error concealment schemes as follows.

Definition 2 (Error Concealment Schemes) *An error concealment scheme is called*

- *EC-1 if it replaces the whole frame (in which some MBs are corrupted) with the previous reconstructed frame;*
- *EC-2 if it replaces a corrupted MB with the MB at the same location from the previous frame;*
- *EC-3 if it replaces the corrupted MB with the MB from the previous frame pointed by a motion vector.*[4]

We would like to stress that EC-1 and EC-2 are special cases of EC-3.

A block diagram of the video decoder is depicted in Fig. 2 where EC-3 is used. Three switches S1, S2, and S3 represent three different scenarios, respectively [14]. Under EC-3, there are three cases for the reconstructed pixel at the receiver as follows.

- *Case (i):* The packet containing F_i^n is received correctly. If F_i^n is intra-coded, then we have $\hat{f}_{ij}^n = \tilde{f}_{ij}^n$, which is illustrated in Fig. 2 with state {S1: Intra, S2: No loss, S3: don't care}. If F_i^n is inter-coded, then we have

  $$\hat{f}_{ij}^n = \hat{e}_{ij}^n + \hat{f}_{uv}^{n-1},$$

  which is illustrated in Fig. 2 with state {S1: Inter, S2: No loss, S3: don't care}.

[3] Packets that arrive later than the maximum expected time are considered lost.

[4] The motion vector of the corrupted MB is copied from one of its neighboring MB when available, otherwise the motion vector is set to zero.

- *Case (ii):* The packet containing F_i^n is lost and the packet containing the MB above ($\bar{F}_i^n$) has been received correctly. Then we have $\hat{f}_{ij}^n = \hat{f}_{ml}^{n-1}$, which is illustrated in Fig. 2 with state {S1: don't care, S2: Loss, S3: MV$\neq$0}.

- *Case (iii):* Both the packet containing F_i^n and the packet containing $\bar{F}_i^n$ are lost. Then we have $\hat{f}_{ij}^n = \hat{f}_{ij}^{n-1}$, which is illustrated in Fig. 2 with state {S1: don't care, S2: Loss, S3: MV=0}.

In the next section, we derive the global distortion metric based on the materials discussed in this section and design an algorithm for optimal mode selection.

3 Optimal Mode Selection

We organize this section as follows. In Section 3.1, we derive the global distortion metrics for an intra-coded MB and an inter-coded MB. In Section 3.2, we design an algorithm for optimal mode selection based on the global distortion metrics.

3.1 Global Distortion Metrics

Without loss of generality, we consider the distortion for macroblock s in frame N, where s ($s \in [0, N_h]$) is the location number and N ($N \geq 0$) is the frame number. Note that the sequence number for both frame and packet start from zero.

Assume the first I-frame of the video stream has been successfully received.[5] Given transition matrix $\mathbf{A}$ for the Gilbert path model, after transmission of n packets, the transition matrix becomes $\underbrace{\mathbf{A} \cdot \mathbf{A} \cdots \mathbf{A}}_{n}$, i.e., $\mathbf{A}^n$, where

$$\mathbf{A}^n = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}^n.$$

Since the resulting $\mathbf{A}^n$ is a 2×2 matrix, we can denote $\mathbf{A}^n$ as

$$\mathbf{A}^n = \begin{bmatrix} P_{LL}^{(n)} & P_{LR}^{(n)} \\ P_{RL}^{(n)} & P_{RR}^{(n)} \end{bmatrix},$$

where $P_{ij}^{(n)}$ ($i \in \{L, R\}$ and $j \in \{L, R\}$) denotes the transition probability from state i to state j after transmission of n packets.

3.1.1 Intra Mode

It has been shown in [14] on how to compute the probability of the event that F_i^n is received correctly and the probability of the event that $\bar{F}_i^n$ is lost. In particular, the probabilities of $\tilde{P}_R^{(i,n)}$ and $\tilde{P}_L^{(i,n)}$ ($n \geq 0$) are given respectively by

$$\tilde{P}_R^{(i,n)} = \begin{cases} 1 & \text{if } X_{\eta_i^n} \in \mathcal{X} \\ P_{RR}^{(1)} & \text{if } \bar{F}_i^n \preceq X_{\eta_i^n} \ (X_{\eta_i^n - 1} \in \mathcal{X} \text{ and } X_{\eta_i^n} \notin \mathcal{X}) \\ P_{RR}^{(\eta_i^n - K + 1)} & \text{if } \bar{F}_i^n \prec X_{\eta_i^n} \ (X_{\eta_i^n} \notin \mathcal{X}) \\ P_{RR}^{(\eta_i^n - K)} \cdot P_{RR}^{(1)} & \text{if } \bar{F}_i^n \preceq X_{\eta_i^n} \ (X_{\eta_i^n} \notin \mathcal{X}) \end{cases},$$

and

$$\tilde{P}_L^{(i,n)} = \begin{cases} 0 & \text{if } X_{\eta_i^n} \in \mathcal{X} \\ P_{RL}^{(1)} & \text{if } \bar{F}_i^n \preceq X_{\eta_i^n} \ (X_{\eta_i^n - 1} \in \mathcal{X} \text{ and } X_{\eta_i^n} \notin \mathcal{X}) \\ P_{RL}^{(\eta_i^n - K + 1)} & \text{if } \bar{F}_i^n \prec X_{\eta_i^n} \ (X_{\eta_i^n} \notin \mathcal{X}) \\ 1 - P_{RR}^{(\eta_i^n - K)} \cdot P_{RR}^{(1)} & \text{if } \bar{F}_i^n \preceq X_{\eta_i^n} \ (X_{\eta_i^n} \notin \mathcal{X}) \end{cases}.$$

It has also been shown in [14] on how to compute the probability of the event that $\bar{F}_i^n$ is received correctly and F_i^n is lost and the probability of the event that both $\bar{F}_i^n$ and F_i^n are lost. Due to paper length constraint, we refer interested readers to [14] for the details of probabilities $\hat{P}_{RL}^{(i,n)}$ and $\hat{P}_{LL}^{(i,n)}$ ($n \geq 0$).

The following proposition shows how to compute MAD for the intra-coded MB under the Gilbert path model and *EC*-3 [14].

Proposition 1 *Under the Gilbert model and EC-3, the MAD for the intra-coded MB at location s of frame N ($N > 0$) is given by*

$$MAD(F_s^N, intra) = \frac{\sum_{j=1}^{256} |f_{sj}^N - E\{\hat{f}_{sj}^N\}|}{256},$$

where

$$E\{\hat{f}_{sj}^n\} = \begin{cases} \tilde{P}_R^{(s,n)} \cdot \tilde{f}_{sj}^n + \tilde{P}_L^{(s,n)} \cdot E\{\hat{f}_{sj}^{n-1}\} & \text{if } F_s^n \in \mathcal{G}^n \\ \tilde{P}_R^{(s,n)} \cdot \tilde{f}_{sj}^n + \hat{P}_{RL}^{(s,n)} \cdot E\{\hat{f}_{ml}^{n-1}\} & \\ \quad + \hat{P}_{LL}^{(s,n)} \cdot E\{\hat{f}_{sj}^{n-1}\} & \text{if } F_s^n \notin \mathcal{G}^n \end{cases}.$$

3.1.2 Inter Mode

The following proposition shows how to compute MAD for the inter-coded MB under the Gilbert path model and *EC*-3 [14].

Proposition 2 *Under the Gilbert model and EC-3, the MAD for the inter-coded MB at location s of frame N ($N > 0$) is given by*

$$MAD(F_s^N, inter) = \frac{\sum_{j=1}^{256} |f_{sj}^N - E\{\hat{f}_{sj}^N\}|}{256},$$

[5] The reason why we make this assumption is to initialize $E\{\hat{f}_{ij}^0\}$ used by the encoding process.

where

$$E\{\hat{f}_{sj}^n\} = \begin{cases} \tilde{P}_R^{(s,n)} \cdot (\tilde{e}_{sj}^n + E\{\hat{f}_{uv}^{n-1}\}) + \tilde{P}_L^{(s,n)} \cdot E\{\hat{f}_{sj}^{n-1}\} \\ \qquad\qquad \text{if } F_s^n \in \mathcal{G}^n \\ \tilde{P}_R^{(s,n)} \cdot (\tilde{e}_{sj}^n + E\{\hat{f}_{uv}^{n-1}\}) + \hat{P}_{RL}^{(s,n)} \cdot E\{\hat{f}_{ml}^{n-1}\} \\ +\hat{P}_{LL}^{(s,n)} \cdot E\{\hat{f}_{sj}^{n-1}\} \\ \qquad\qquad \text{if } F_s^n \notin \mathcal{G}^n \end{cases}$$

We would like to stress that Proposition 2 holds for any packetization scheme.

Although the global distortion metrics derived in Section 3.1 only apply to the Gilbert path model, the methodology we employ (i.e., the end-to-end approach) is general and can be applied to any path model (e.g., self-similar path model).

3.2 Globally Optimized Mode Selection

Given the packetization scheme used by the source, the path characteristics and the error concealment scheme used by the decoder, we design a globally R-D optimized mode selection algorithm.

Consider a GOB denoted by $\mathcal{F}_g^n = (F_g^n, \cdots, F_{g+N_G-1}^n)$, where N_G is the number of MBs in a GOB. Assume each MB in $\mathcal{F}_g^n$ can be coded using only one of the two modes in set $\mathcal{I}$. Then for a given GOB, the modes assigned to the MBs in $\mathcal{F}_g^n$ are given by the N_G-tuple, $\mathcal{M}_g^n = (M_g^n, \cdots, M_{g+N_G-1}^n) \in \mathcal{I}^{N_G}$. The problem of globally R-D optimized mode selection is to find the combination of modes that minimizes the distortion for a given GOB, subject to a constraint R_c on the number of bits used. This constrained problem can be formulated as

$$\min_{\mathcal{M}_g^n} D(\mathcal{F}_g^n, \mathcal{M}_g^n) \quad \text{subject to} \quad R(\mathcal{F}_g^n, \mathcal{M}_g^n) \le R_c, \tag{2}$$

where $D(\mathcal{F}_g^n, \mathcal{M}_g^n)$ and $R(\mathcal{F}_g^n, \mathcal{M}_g^n)$ denote the total distortion and bit budget, respectively, for the GOB $\mathcal{F}_g^n$ with a particular mode combination $\mathcal{M}_g^n$.

The constrained minimization problem in (2) can be converted to an unconstrained minimization problem by Lagrange multiplier technique. Under the assumption of an additive distortion measure, the Lagrangian cost function can be given by

$$\begin{aligned} J(\mathcal{F}_g^n, \mathcal{M}_g^n) &= \sum_{i=g}^{g+N_G-1} J(F_i^n, \mathcal{M}_g^n) \\ &= \sum_{i=g}^{g+N_G-1} [D(F_i^n, \mathcal{M}_g^n) + \lambda R(F_i^n, \mathcal{M}_g^n)]. \end{aligned}$$

Thus, the objective function becomes

$$\min_{\mathcal{M}_g^n} \left\{ \sum_{i=g}^{g+N_G-1} J(F_i^n, \mathcal{M}_g^n) \right\}. \tag{3}$$

If both the rate and distortion for macroblock F_i^n are not affected by other mode that is not used by macroblock F_i^n, a simplified Lagrangian can be given by

$$J(F_i^n, \mathcal{M}_g^n) = J(F_i^n, M_i^n).$$

Thus, the optimization problem of (3) reduces to

$$\sum_{i=g}^{g+N_G-1} \min_{M_i^n} J(F_i^n, M_i^n) = \sum_{i=g}^{g+N_G-1} \min_{M_i^n} \{D(F_i^n, M_i^n) + \lambda R(F_i^n, M_i^n)\}, \tag{4}$$

where the global distortion $D(F_i^n, M_i^n)$ can be expressed by the formulae we derived in Section 3.1, according to the coding mode, the packetization scheme used by the source and the error concealment scheme used by the decoder.

The problem of (4) is a standard R-D optimization problem and can be solved by the approaches described in [4, 6, 9, 12]. Different from these approaches, we use a simpler method to obtain λ.

Since a large λ in the optimization problem of (4) can reduce the bit-count of the coded frame, we employ this nature in choosing λ. To be specific, at the end of frame n, we adjust λ for frame $n+1$ (i.e., λ_{n+1}) as follows:

$$\lambda_{n+1} = \frac{2 \cdot B_n + (\gamma - B_n)}{B_n + 2 \cdot (\gamma - B_n)} \cdot \lambda_n, \tag{5}$$

where B_n is the current buffer occupancy at the end of frame n and γ is the buffer size. λ_n is initialized by a preset value λ_0. The adjustment in Eq. (5) is to keep the buffer occupancy at the middle level to reduce buffer overflow or underflow. In other words, Eq. (5) also achieves the objective of rate control.

Sections 2 and 3 complete the theoretical part of our work. To evaluate the effectiveness of our approach, we apply our theory to a specific system – an architecture for point-to-point MPEG-4 video conferencing over the Internet.

4 An Application for MPEG-4 Video

In this section, we present an end-to-end architecture for point-to-point MPEG-4 video conferencing over the Internet.

4.1 Architecture

Figure 3 shows our end-to-end architecture for point-to-point MPEG-4 video conferencing over the Internet. We use the MPEG-4 rate control algorithm described in [2, 13] to control the output rate to be constant. In this paper, we set the rate fixed to investigate the error resilient capability of our architecture and algorithm.

In Fig. 3, on the sender side, raw bit-stream of live video is encoded by an MPEG-4 encoder. After this stage, the compressed video bit-stream is first packetized at the sync layer and then passed through the RTP/UDP/IP layers before entering the Internet. Packets may be dropped at a router/switch (due to congestion) or at the destination (due to excess delay). For packets that are successfully delivered to the destination, they first pass through the RTP/UDP/IP layers in reverse order before being decoded at the MPEG-4 decoder.

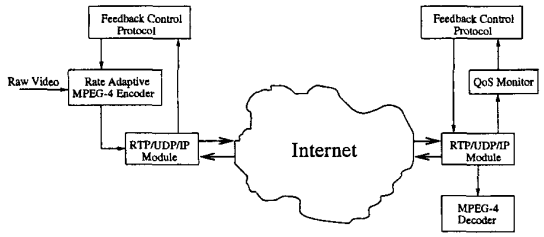

Figure 3: An end-to-end architecture for MPEG-4 video conferencing.

On the receiver side, a QoS monitor detects the packet loss through RTP sequence number and estimates the transition probabilities (e.g., p and q in the Gilbert model). When the receiver sends out a feedback RTCP packet to the source, the estimated parameters p and q are carried in the feedback packet. Once source receives such feedback, it encodes the video based on the parameters p and q through the proposed R-D optimized mode selection in Section 4.4 and rate control algorithm in [2, 13].

4.2 Feedback Control Mechanism

Because UDP does not guarantee packet delivery, the receiver needs to rely on upper layer (i.e., RTP/RTCP) to detect packet loss [7]. RTCP provides QoS feedback through the use of *Sender Reports (SR)* and *Receiver Reports (RR)* at the source and destination, respectively. The feedback control protocol employs RTCP to convey QoS information so that QoS information can be utilized by the encoder. Packet loss can be detected by the QoS monitor through examining the RTP packet sequence number at the receiver side. On the other hand, a packet that arrives after the maximum delay threshold is considered lost.

Since we employ the Gilbert path model, the path characteristics can be estimated by the QoS monitor as follows. Upon obtaining the packet loss information from RTP/UDP/IP module, the QoS monitor measures N_1, N_2, N_3, and N_4, and estimates the transition probabilities, p and q, through Eq. (1).

The period for estimating p and q is set to five seconds. That is, N_1, N_2, N_3, and N_4 are measured during the five-second period. At the end of each period, p and q are obtained through (1) and then conveyed to the source through an RTCP packet. Since N_1, N_2, N_3, and N_4 are reset to zero at the end of each period, the estimated p and q reflect the current network congestion status.

4.3 Packetization and Error Concealment

We use PKT-3 rather than PKT-2 to achieve efficiency for Internet video conferencing. In addition, when a packet is lost, we employ EC-3 to conceal the region associated with the lost packet. To be specific, each corrupted MB will be replaced with the MB in the previous frame pointed by an estimated motion vector. The estimated motion vector of the corrupted MB is copied from the MB above it when available, otherwise the motion vector is set to zero. Note that a more sophisticated error concealment scheme than EC-3 may achieve better performance than that of EC-3.

4.4 Feedback-Based Optimal Mode Selection

For implementation purpose, our end-to-end approach also considers the impact of feedback mechanism on the video quality (in terms of global distortion). The rationale is as follows. Global optimality is not achievable without feedback since the source could not select an optimal mode without knowledge of the path characteristics and receiver behavior. In addition, the congestion status of the Internet is dynamically changing. Assigning the path characteristics (e.g., p and q) with fixed numbers may either lose compression efficiency when the network is less congested than expected, or suffer from insufficiency of error resilience when the network becomes heavily congested. Therefore, it is not valid to assume that the path characteristics is known *a priori* and is fixed for the the real Internet. From our experiments and simulations, we observe that the percentage of intra-coded macroblocks should increase as the packet loss ratio increases in order to improve the capability of error resilience. Thus, MPEG-4 video coding should adapt to the changing Internet environment, i.e., network congestion.

This motivates us to employ a feedback mechanism to convey such information to the encoder as the path characteristics (i.e., p and q) estimated at the receiver and the error concealment scheme used by the decoder. The type of error concealment scheme used by the receiver can be transmitted at the set-up period of the session.

Base on the theory in Section 3, we design a globally R-D optimized mode selection algorithm for MPEG-4. Since we employ PKT-3, EC-3 and the Gilbert path model, both the global distortion $D(F_i^n, intra)$ and the global distortion $D(F_i^n, inter)$ (see 4) can be calculated using results in previous sections. Based on the feedback p and q, the proposed algorithm will choose a mode which has the best trades-off between compression efficiency and error resilience.

5 Simulation Results

In this section, we implement the end-to-end architecture described in Section 4 on our network simulator and perform a simulation study of video conferencing with MPEG-4.

5.1 Simulation Settings

The network configuration that we use is the *chain* network (Fig. 4). In our simulations, path G1 consists of one MPEG-4 source, three TCP connections and three UDP connections while paths G2, G3 and G4 all consist of three TCP connections and three UDP connections, respectively. The link capacities on Link12, Link23, and Link34 are all 350 Kbps.

We implement the architecture depicted in Fig. 3 for MPEG-4 video. At the source side, we use the standard raw video sequence "Miss America" in QCIF format for the video encoder. The encoder employs the rate control described in [2, 13] to keep a constant rate at 100 Kbits/s. The frame rate is 10 frames/s. The encoder is used in the rectangular mode, with intra-VOP refreshment period of 50 frames.

The encoded bit-stream is packetized with PKT-3 scheme (i.e., a packet corresponds to a GOB). Additional

Table 2: Simulation parameters.

End system	MPEG-4	MaxPL	526 bytes
		Rate	100 Kbps
		Frame rate	10 frames/s
		I-VOP refreshment period	50 frames
		λ_0	1
		Buffer size	1 Mbytes
	TCP	Mean packet processing delay	300 μs
		Packet processing delay variation	10 μs
		Packet size	576 bytes
		Maximum receiver window size	64K bytes
		Default timeout	500 ms
		Timer granularity	500 ms
		TCP version	Reno
	UDP	$E(T_{on})$	100 ms
		$E(T_{off})$	150 ms
		r_p	100 Kbps
		Packet size	576 bytes
Switch		Buffer size	10 Kbytes
		Packet processing delay	4 μs
Link	End system to switch	Link speed	10 Mbps
		Distance	1 km
	Switch to switch	Distance	1000 km

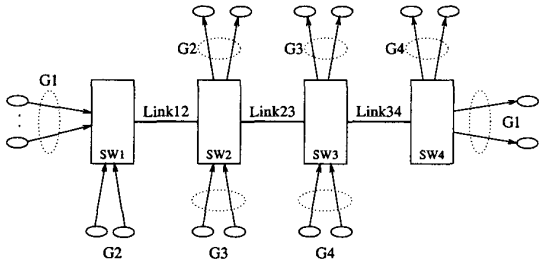

Figure 4: A chain network.

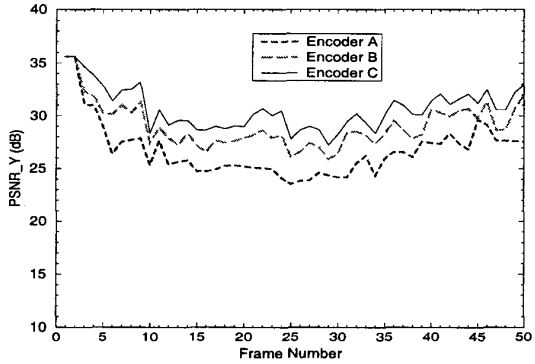

Figure 5: PSNR at the receiver under the chain network.

overhead from RTP/UDP/IP is also added to the packet before it is sent to the network. We use 576 bytes for the path MTU. Therefore, the maximum payload length, MaxPL, for MPEG-4 is 526 bytes (576 bytes minus 50 bytes of overhead) [8]. Packets may be dropped due to congestion in the network. For arriving packets, the receiver extracts the packet content to form the bit-stream for the decoder.

In addition to the MPEG-4 video, we also use TCP/UDP connections to simulate the background interfering traffic. All TCP sources are assumed to be persistent during the simulation run. For UDP connections, we use an exponentially distributed on/off model with average $E(T_{on})$ and $E(T_{off})$ for on and off periods, respectively. During each on period, the packets are generated at peak rate r_p. The average bit rate for a UDP connection is, therefore, $r_p \cdot \frac{E(T_{on})}{E(T_{on})+E(T_{off})}$.

Table 2 lists the parameters used in our simulation.

Under such simulation settings, we consider three different encoders for MPEG-4 video as follows.

Encoder A: employs the classical approach for R-D optimized mode selection.

Encoder B: implements the globally R-D optimized mode selection described in Section 4.4, *except* that the feedback mechanism is disabled.

Encoder C: implements the globally R-D optimized mode selection described in Section 4.4. Here, the feedback mechanism is employed.

5.2 Results

We run our simulation for 100 seconds. Since there are only 150 continuous frames in "Miss America" sequence available, we repeated the video sequence cyclically during the simulation run.

Figure 5 shows the PSNR of the first 50 frames in the 100-second simulations. We observed that Encoder C achieves the best performance, Encoder B has the second best performance, and Encoder A performs the worst. That is, our approach achieves better performance than the classical one, even if feedback mechanism is not employed; feedback-based scheme (i.e., Encoder C) achieves better performance than non-feedback-based

scheme (i.e., Encoder B).

For the 100-second simulation run, the average PSNRs for Encoders A, B, and C are 26.9 dB, 28.9 dB and 30.4 dB, respectively. The packet loss ratios are all 3.2%.

6 Concluding Remarks

In this paper, we investigated the fundamental problem of R-D optimized mode selection for point-to-point Internet video communication from an end-to-end perspective, which includes source behavior, path characteristics, and receiver behavior. We introduced a notion of *global* distortion by taking into consideration of both the path characteristics and the receiver behavior, in addition to the source behavior. We derived, for the first time, a set of accurate global distortion metrics for any packetization scheme. Equipped with the global distortion metrics, we designed an R-D optimized mode selection algorithm to provide the best trade-off between compression efficiency and error resilience. As an application, we applied our theory to point-to-point MPEG-4 video conferencing over the Internet. Simulation results conclusively demonstrated that our end-to-end approach offers superior performance over the classical approach for Internet video conferencing.

References

[1] J-C. Bolot and T. Turletti, "Adaptive error control for packet video in the Internet," in *Proc. IEEE ICIP'96,* Sept. 1996.

[2] T. Chiang and Y.-Q. Zhang, "A new rate control scheme using quadratic rate distortion model," *IEEE Trans. on Circuits and Systems for Video Technology,* vol. 7, no. 1, pp. 246–250, Feb. 1997.

[3] ISO/IEC JTC 1/SC 29/WG 11, "Information technology - coding of audio-visual objects, part 1: systems, part 2: visual, part 3: audio," FCD 14496, Dec. 1998.

[4] J. Lee and B. W. Dickinson, "Rate-distortion optimized frame type selection for MPEG encoding," *IEEE Trans. on Circuits and Systems for Video Technology,* vol. 7, no. 3, pp. 501–510, June 1997.

[5] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Magazine,* pp. 74–90, Nov. 1998.

[6] K. Ramchandran, A. Ortega and M. Vetterli, "Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders," *IEEE Trans. on Image Processing,* vol. 37, pp. 533–545, Aug. 1994.

[7] H. Schulzrinne, S. Casner, R. Frederick and V. Jacobson, "RTP: a transport protocol for real-time applications," *RFC 1889,* Internet Engineering Task Force, Jan. 1996.

[8] H. Schulzrinne, D. Hoffman, M. Speer, R. Civanlar, A. Basso, V. Balabanian and C. Herpel, "RTP payload format for MPEG-4 elementary streams," *Internet Draft,* Internet Engineering Task Force, March 1998.

[9] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. on Acoust., Speech and Signal Processing,* vol. 36, no. 9, pp. 1445–1453, Sept. 1988.

[10] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine,* pp. 74–90, Nov. 1998.

[11] T. Turletti and C. Huitema, "RTP payload format for H.261 video streams," *RFC 2032,* Internet Engineering Task Force, Oct. 1996.

[12] T. Wiegand, M. Lightstone, D. Mukherjee, T. G. Campbell and S. K. Mitra, "Rate-distortion optimized mode selection for very low bit rate video coding and the emerging H.263 standard," *IEEE Trans. on Circuits and Systems for Video Technology,* vol. 6, no. 2, pp. 182–190, April 1996.

[13] D. Wu, Y. T. Hou, W. Zhu, H.-J. Lee, T. Chiang, Y.-Q. Zhang and H. J. Chao, "On end-to-end architecture for transporting MPEG-4 video over the Internet," to appear in *IEEE Trans. on Circuits and Systems for Video Technology,* 2000.

[14] 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," to appear in *IEEE J. on Select. Areas in Commun.,* 2000.

[15] C. Zhu, "RTP payload format for H.263 video streams," *RFC 2190,* Internet Engineering Task Force, Sept. 1997.