Network-Adaptive Low-Latency Video Communication Over Best-Effort Networks

Yi J. Liang, *Member, IEEE,* and Bernd Girod, *Fellow, IEEE*

***Abstract*—The quality of service limitation of today's best-effort networks poses major challenge for low-latency video communication. To combat network losses for real-time and on-demand video communication, which exhibits stronger dependency across packets, a network-adaptive coding scheme is employed to dynamically manage the packet dependency using optimal reference picture selection. The selection of the reference is achieved within a rate-distortion optimization framework and is adapted to the varying network conditions. For network-adaptive streaming of prestored video, based on an accurate loss-distortion model, a prescient scheme that optimizes the dependency of a group of packets is proposed to achieve global optimality as well as improved rate-distortion performance. With the improved trade-off between compression efficiency and error resilience, the proposed system does not require retransmission of lost packets, which makes less than one-second low-latency communication possible.**

***Index Terms*—Error resilience, H.264, low latency, network-adaptive video coding, rate-distortion optimization, reference picture selection, video streaming.**

I. Introduction

SINCE the introduction of the first commercial products in 1995, Internet video communication has experienced phenomenal growth [1]. However, despite of the rapid expansion of the underlying infrastructure, technological challenges are still a major barrier to the wide adoption of online streaming media today. Internet video communication today is plagued by variability in throughput, packet loss, and delay, due to network congestion and the heterogeneous infrastructure. To mitigate these effects, media streaming systems typically employ a large receiver buffer that introduces a latency of 5–15 s. This is undesirable since the slow start-up is annoying and high latency severely impairs the interactive playback features, such as VCR functionality.

In contrast, for IP-based speech communication, the end-to-end latency can usually be kept on the order of a hundred milliseconds [2]. Nevertheless, low latency in video streaming is much more difficult due to the sensitivity of the compressed video stream against channel losses. For speech coding, the dependency across successive data units is weak or there is no dependency at all. For typical motion-compensated video coding, as is used in most of today's codecs, this dependency is much stronger. An inter-coded frame is predicted from a reference picture with motion compensation, so that the temporal redundancy across successive pictures is removed or reduced to provide higher coding efficiency. However, proper decoding of such inter-coded pictures depends on the error-free reception and reconstruction of the reference picture it uses, which is not guaranteed over lossy networks.

Assume a simplified scenario where an IP packet contains one video frame. If a packet (frame) is lost, the proper reconstruction of all subsequent frames that depend on the lost frame is affected. Hence, in a typical automatic repeat request (ARQ)-based system, whenever a packet is lost, retransmission is required to guarantee the correct reception of each frame. However, the time for retransmission constitutes the major part of the undesirable end-to-end delay.

In this paper, we are to address the latency issue in real-time and streaming video applications that demand very low communication delay. Here we are not focusing on video applications that can torlerate latency of up to a few seconds, but aiming to achieve latency of only hundreds of miliseconds. For conversational real-time applications, low latency is extreamly desirable since excess delay impaires communication interactivity. Even for on-demand streaming applications in which delay requirements used to be considered as more relaxed, voice over IP (VoIP)-like low latency of hundreds of miliseconds significantly improves interactive playback features, such as random indexing, fast-forwarding and switching channels, which will, hence, fundamentaly change the typical user experiences today.

To address the various challenges for video communication, research efforts in recent years have particularly been directed toward communication efficiency, error-robustness, and low latency [3]–[10]. Many of the recent algorithms use rate-distortion (R-D) optimization techniques to improve the compression efficiency [11]–[13], as well as to increase the error-resilient performance over lossy networks [14], [15]. The goal of these optimization algorithms is to minimize the expected distortion due to both compression and channel losses subject to the bit-rate constraint.

As is mentioned above, ARQ techniques incorporate channel feedback and employ the retransmission of erroneous data [16]–[20]. ARQ intrinsically adapts to the varying channel conditions and tends to be more efficient in transmission. However, for real-time communication and low-latency streaming, the latency introduced by ARQ is a major concern.

Examples of different error-resilience schemes that introduce lower latency than ARQ include intra/inter-mode switching

Manuscript received January 7, 2004; revised October 15, 2004. This work was completed at the Department of Electrical Engineering, Stanford University, and was supported in part by Hewlett Packard Laboratories and in part by the Stanford Network Research Center (SNRC). This paper was recommended by J. Arnold.

Y. J. Liang is with Qualcomm CDMA Technologies, San Diego, CA 92121 USA (e-mail: yiliang@stanfordalumni.org).

B. Girod is with the Department of Electrical Engineering, Stanford University, CA 94305 USA.

Digital Object Identifier 10.1109/TCSVT.2005.856919

[21]–[24], where intra-coded macroblocks are updated according to the network condition to mitigate temporal error propagation, and forward error correction (FEC), in which case missing packets can be recovered at the receiver as long as a sufficient number of packets is received [25]–[29]. In many cases over IP and wireless networks, however, the loss across successive packets is correlated. A packet loss may often be followed by a burst of losses, which significantly decreases the efficiency of FEC schemes. In order to combat burst loss, redundant information has to be added into temporally distant packets, which introduces higher delay at both encoder and decoder sides. Hence, the repair capability of FEC is limited by the delay budget in very-low-latency applications.

Another approach is to modify the temporal prediction dependency of motion-compensated video coding in order to mitigate or stop error propagation. Example implementations include reference picture selection (RPS) [3], [30]–[32] and NEWPRED in MPEG-4 [33], where channel feedback is utilized and new references may be selected to efficiently stop error propagation due to any transmission error. More recently, dynamic control of the prediction dependency has been presented using long-term memory (LTM) prediction to achieve improved R-D performance [13], [14], [34]. Another scheme is the video redundancy coding (VRC), where the video sequence is coded into independent threads in a round-robin fashion to achieve better resilience against errors [3].

Recently, the schemes of packet transmission scheduling have also been proposed to optimally allocate the time and bandwidth resources among packets and decide which packet to transmit or retransmit at each opportunity, in order to achieve improved error-resilience under the rate constraint [29], [35]–[39].

The error-resilience techniques discussed above are employed to improve the overall system performance at different data rate and delay costs. They provide improved error resilience through source/network coding and optimizing the transmission policy, the amount of redundancy in an FEC system, and the prediction dependency across packets.

The major contribution of this paper is the development of error-resilient coding techniques that achieve *very* low latency for real-time and on-demand video applications that impose highly stringent delay requirements. In such scenarios, packets are transmitted as soon as they become available at the sender, and retransmissions cannot be afforded. In order to increase error resilience and eliminate the need for retransmission, we account for the prediction dependency across packets resulting from hybrid video coding, and dynamically manage this dependency to achieve increased error resilience at a given data rate constraint. The increased error resilience eliminates the need of retransmitting lost packets, which enables less than one-second low-latency video streaming. The proposed scheme applies to both live and prestored video streaming and is compatible with open standards such as ISO/IEC MPEG-4 [33] and ITU-T H.264 [32].

This paper is organized as follows. We first introduce the concept of packet dependency management and its implementation. In Section III, we present the algorithm of channel-adaptive optimal packet dependency management. Following that we address the issues for the streaming of pre-encoded video in Section IV, where a prescient scheme of dependency management for a group of packets is presented to achieve further improved performance. Experimental results are presented in Section V.

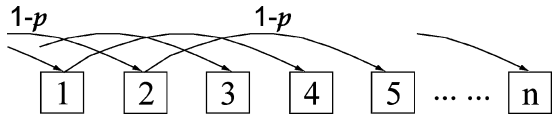


Fig. 1. Coding structure where each frame uses the third previous frame as a reference ($v = 3$). Each frame is correctly received at the decoder with probability $1 - p$. Frame 5 in the sequence depends on $\lceil 5/3 \rceil = 2$ previous frames, and the probability it will be affected by a previous loss is $p_e = 1 - (1 - p)^2$.

II. Packet Dependency Management and Reference Picture Selection

In manipulating the prediction dependency, an earlier related proposal is the RPS mode in Annex N of H.263+ to terminate error propagation based on feedback [3], [31]. When the encoder learns through the feedback channel that a previous frame is lost, instead of using the most recent frame as a reference, it can code the next P-frame based on an older frame that is known to be correctly received at the decoder [40], [41]. The multiframe prediction support in Annex N was later subsumed by the more advanced Annex U of H.263++ and is now an integral part of the emerging H.264 standard [32]. In our work [42], we extend the RPS concept by allowing the use of a reference frame whose reception status is uncertain but whose reliability can be inferred, for live-encoding.

In [14] and [34] LTM prediction is used for both improved coding efficiency and error resilience over wireless networks. Different macroblocks in a frame may be predicted from different reference frames, which makes it difficult to put an entire frame into an IP packet and manage the prediction dependency at the packet level during transmission. Throughout this work, we select the reference at the frame level and assume that each predictively coded frame is coded into one IP packet (which can be extended to more general cases). In this way we manage the frame prediction dependency at the packet level.

In a conventional encoding and transmission scheme without any awareness of network losses, an I-frame is typically followed by a series of P-frames, which are predicted from their immediate predecessors. This scheme is vulnerable to network errors since each P-frame depends on its predecessor and any packet loss will break the prediction chain and affect all subsequent P-frames. If each P-frame is predicted from the frame preceding the previous frame instead, the scheme is more robust against network errors due to the changed dependency and the higher certainty of the reference frame. Consider, for example, a fixed coding structure where each frame uses the reference that is v frames back for prediction, where v is used to denote the *coding mode*, or *prediction mode*. The nth frame in the sequence thus depends on $\lceil n/v \rceil$ previous frames, where $\lceil x \rceil$ represents the smallest integer number that is greater than or equal to x. An example for $v = 3$ is illustrated in Fig. 1. Assuming each packet is lost independently with probability p, the probability that the

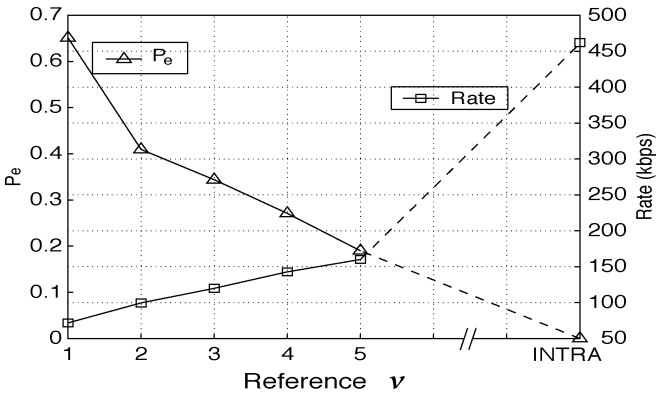


Fig. 2. Probability of the 10th frame being affected by a prior loss (left axis) and the sequence-averaged rates (right axis) using different reference frames. Rates are obtained by encoding the first 230 frames of *Foreman* sequence (30 frame/s) using H.264 TML 8.5 at an average PSNR of approximately 33.4 dB. $p = 0.10$.

nth frame in the sequence will be affected by a previous loss is hence

$$p_e = 1 - (1-p)^{\lceil \frac{n}{v} \rceil}. \tag{1}$$

This probability is plotted in Fig. 2 for $p = 0.10$, $n = 10$, and $v = 1, 2, \ldots, 5$, and intra-coding (we use $v = \infty$ to denote intra-coding).

As illustrated in Fig. 2, using frames from the LTM with $v > 1$ for prediction, instead of using an immediately previous frame $(v = 1)$, reduces prediction efficiency and increases error resilience. The robustness is normally obtained at the expense of a higher bitrate since the correlation between two frames becomes weaker in general as they are more widely separated. A special and extreme case is the I-frame, which is the most robust over lossy networks, but generally requires 5–10 times as many bits as the P-frame. In Fig. 2, we also show the average rates of encoding the *Foreman* sequence at close peak signal-to-noise ratios (PSNRs) using different coding modes v, including intra-coding. More advanced coding modes achieve higher error resilience at the cost of higher bitrate. In practice, especially in a rate-constrained environment, this type of technique is to be used in conjunction with rate control techniques to allow the media stream to be network friendly.

Fixed reference selection schemes provide different amount of error resilience at different coding costs, as is shown in Fig. 2. In this work, we consider the dependency across packets and dynamically manage this dependency while adapting to the varying network conditions. Due to the trade-off between error resilience and coding efficiency, we apply *optimal reference picture selection (ORPS)* within an R-D framework, by considering video content, channel loss probability and channel feedback.

III. R-D Optimized Network-Adaptive Packet Dependency Management

We take the network loss into consideration and select the reference picture within an R-D framework. We use a binary tree structure to represent error propagation from frame to frame and all possible decoded outcomes at the decoder.

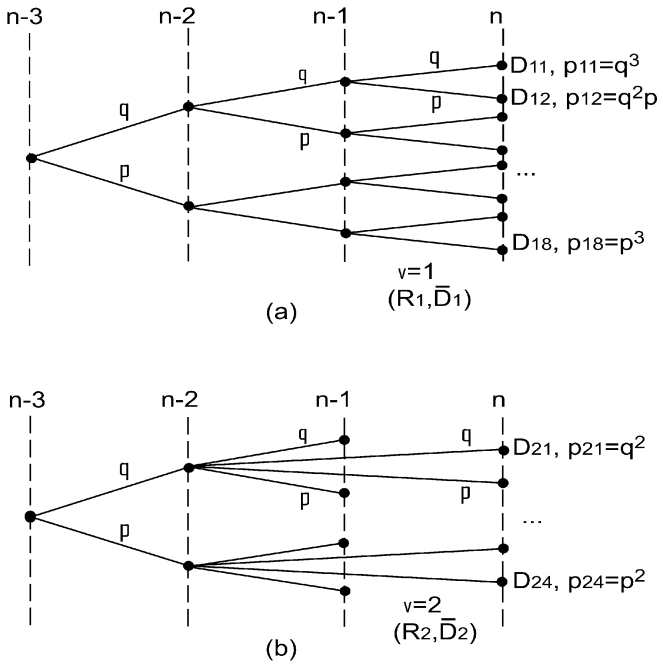


Fig. 3. Binary tree structure for estimating error propagation and optimal reference selection. Frames $n-2$ and $n-1$ have already been coded. When coding Frame n, plot (a) depicts the case using Frame $n-1$ as the reference for prediction $(v = 1)$. Plot (b) depicts the case using Frame $n-2$ as the reference for prediction $(v = 2)$.

Fig. 3 illustrates such a tree structure. A *node* in the tree represents a possible decoded outcome (frame) at the decoder. In the example shown in Fig. 3, Frame $n-3$ has only one node with probability 1 (e.g., due to the reception status confirmed by feedback). Frames $n-2$ and $n-1$ both, for instance, use their immediately preceding frames as references. Two *branches* leave the node of Frame $n-3$ representing two cases that either reference frame $n-3$ is properly received (and decoded) with probability q or lost with probability $p = 1-q$. These two cases lead to two different versions of Frame $n-2$, provided that Frame $n-2$ is available at the decoder. The upper node of Frame $n-2$ is obtained by normal decoding process using the correct reference (decoded $n-3$); and the lower node corresponds to the case when Frame $n-3$ is lost. In the latter case, a simple concealment is done by copying $n-4$ to $n-3$, and Frame $n-2$, hence, has to be decoded using the concealed reference (decoded $n-4$). This leads to the mismatch error that might propagate at the decoder, depending on the prediction dependency of the following frames. The distortion associated with these two cases are evaluated by decoding $n-2$ at the *encoder* side. The value of q or p is estimated from the accumulated channel statistics, which may be updated as the network conditions change.

In encoding the next frame n, several trials are made using different reference frames in the LTM and the resulting rates and expected distortion are obtained in order to calculate the cost function. The costs associated with all candidate coding modes are compared and the best reference is selected that achieves the minimum cost.

Assuming V previously decoded frames are available from the LTM (V is referred to as the *length of LTM*), we use $v(n)$ to represent the reference frame, or the coding mode, that Frame

n may use. For example $v(n) = 1$ denotes using the previous frame and $v(n) = 2$ denotes using the frame preceding that frame, and so on. For a particular $v(n) = v$, a rate R_v is obtained and the expected distortion over all decoded outcomes is given by

$$\overline{D}_v = \sum_{l=1}^{L(n)} p_{vl} D_{vl} \tag{2}$$

where $L(n)$ is the number of nodes for Frame n, and $L(n) = 2L(n - v(n))$. p_{vl} is the probability of outcome (node) l, which can be calculated easily from the tree structure, if statistical independence of successive losses is assumed. For example, in Fig. 3, $p_{11} = q^3$, while $p_{12} = q^2 p$ and so on. D_{vl} is the distortion associated with the decoded outcome l. Note that D_{vl} includes both the quantization error and possible decoding mismatch error, which is calculated accurately at the encoder side. For simplicity we assume one frame is contained in one IP packet in this work. If a frame is large enough and split into more than one packet, the frame can still be represented by a node in this model, but the outcome probabilities will slightly change in the tree structure.

With the obtained rate and expected distortion, the Lagrangian cost associated with using the reference frame $v(n) = v$ is

$$J_v = \overline{D}_v + \lambda R_v. \tag{3}$$

Comparing all candidate reference frames, $v(n) = 1, 2, \ldots, V$ and ∞ (intra-coding), the optimal reference frame $v_{\text{opt}}(n)$ is the one that results in minimal J_v

$$v_{\text{opt}}(n) = \arg\min_{v=1,2,\ldots,V,\infty} J_v(n). \tag{4}$$

In (3), λ is a Lagrange multiplier and we use $\lambda = 5e^{0.1Q}((5 + Q)/(34 - Q))$, which is the same as λ_{mode} in H.264 TML 8 used to select the optimal prediction mode [43], [44], and Q is the quantization parameter used to trade off rate and distortion. Note that in each stage $v_{\text{opt}}(n)$ is obtained given the condition that Frame n is available at the decoder, which means the reception status of n is not considered in selecting the reference frame for n at the encoder.

In [14], a similar binary tree structure is used to select the optimal reference for each macroblock. However only three branches are considered in the binary tree to calculate the approximate expected distortion, and a second Lagrange multiplier κ is used to trade off error resilience and coding efficiency. An approximate model has to be used in [14] because of several difficulties mentioned: 1) tree size grows at the rate of $L = 2^n$ for each macroblock and modeling can be soon intractable and 2) time instants do not correspond to levels of the tree due to LTM prediction.

In this work, we take the advantage of network feedback in order to limit the tree size, which is reasonable and necessary in practice. When the reception status is known (e.g., from feedback) for a previously sent packet, half of the branches leaving the corresponding node are removed, and so are all the dependant nodes. In this way, the maximum size of the tree is kept constant as the states propagate. In general, given the maximal feedback delay d_{fb} in frames (when encoding Frame n, the status of Frame $n - d_{\text{fb}}$ becomes known), the maximum number of outcomes kept for a frame is $L = 2^{d_{\text{fb}}-1}$. The feedback may be obtained through different means of transport used in the system. For example, from the ACK flag or time-out in ARQ-based systems (although retransmission is not needed for the scheme), or the RTCP in general IP systems. Note that the reference selection algorithm itself does not necessarily depend on any feedback. In a practical system, the feedback does not have to be prompt either, although faster feedback helps to improve the performance.

In solving the second difficulty mentioned above, we keep tracking the states of each frame by storing all possible decoded outcomes in the LTM of the encoder, so that time instants can be tracked and represented by the levels of the tree. For each frame to be encoded, all of its possible decoding outcomes are obtained using the saved outcomes of the reference frame (either correct or concealed) from the LTM. Therefore, the binary tree modeling here is accurate in estimating distortions. This is achieved at higher storage complexity, depending on the length of LTM, V, and the feedback delay d_{fb}. Noting that the maximum decoded frames kept for a most recent frame encoded is $L = 2^{d_{\text{fb}}-1}$, the maximum number of decoded frames that have to be stored in the LTM is

$$\begin{cases} \sum_{k=d_{\text{fb}}-1-V}^{d_{\text{fb}}-1} 2^k, & \text{for } V \le d_{\text{fb}} - 1 \\ [V - (d_{\text{fb}} - 1)] + \sum_{k=0}^{d_{\text{fb}}-1} 2^k, & \text{for } V > d_{\text{fb}} - 1. \end{cases} \tag{5}$$

For instance, when using $V = 5$ and $d_{\text{fb}} = 7$, the maximum number of decoded frames that have to be stored is 126, which corresponds to about 4.6 MB $(1.5 \times 176 \times 144 \times 126/1024^2)$ for QCIF sequences. The increased memory overhead is affordable and worthwhile for the media server when considering the gain in error resilience and low latency.

IV. Prescient Packet Dependency Management for Pre-Encoded Media

For live applications, video is compressed and coded on-the-fly, which allows network-adaptive coding to be implemented using the scheme presented in the last section. For on-demand streaming, in coding prestored video, we can also take advantage of precomputed information to further optimize the performance at service. In Section III, we optimize the prediction dependency of the *next* packet to transmit for live-encoding. The optimality achieved with this *greedy* method is local. For streaming of pre-encoded media, we optimize the prediction dependencies for a *group* of packets before they are transmitted, by using the precomputed R-D information for precompressed streams. This approach, which we refer to as the *prescient* method, may achieve global R-D optimality for a group of packets, or even the entire sequence. The scheme can be realized at low complexity in practical systems. A prescient scheme for selecting the optimal intra/inter coding mode for macroblocks was previously proposed in [45], which achieves optimality for current and a limited number (one or two) of subsequent frames. In this section, we consider a larger group that contains more frames in determining the optimal prediction dependency.

A. Prescient Packet Dependency Management

For streaming of precompressed video, we take advantage of precomputed rate and distortion information in the R-D preamble, which is actually a compact description of packet contents. With the auxiliary information computed and stored offline, we are able to determine the R-D optimized coding modes by only considering the R-D preamble instead of actually compressing the video, which greatly reduces the complexity at streaming time [46].

Unlike the greedy method in Section III, we predetermine the coding modes for a *group* of frames before they are sent. This group of frames can be a GOP starting with a traditional I-frame, or an instantaneous decoder refresh (IDR) picture proposed in H.264 [32], which resets the multiframe buffer to break the inter-dependencies from any picture decoded prior to the IDR picture. In this case the group is independent of previous groups sent if its leading frame is properly decoded. The group of frames can also be a sub-GOP as we proposed in [47], and the leading frame of the group is predicted from the leading frame of the last sub-GOP. In this case, all the frames except the leading frame of the group are independent of previous groups.

For a group of L frames indexed by $i = 0, 1, \ldots, L-1$, we use $\mathbf{v} = (v_0, v_1, \ldots, v_{L-1})$ to denote their prediction modes. For the leading frame of the group, $i = 0$, we restrict its coding mode to be either intra or $v_0 = L$, and the mode to use is determined by examining respective R-D costs, using any feedback information available. The ith frame following the leading frame is allowed to use coding modes $1 \leq v_i \leq i$ and intra, e.g., it is only allowed to be predicted from prior frames in the same group. This keeps different groups relatively independent of each other, and helps to avoid potential mismatch error that might occur when assembling and transmitting pre-encoded bitstreams [47]. The coding modes of the frames following the leading frame are determined in a prescient fashion. We use $R_{\mathbf{v}}$ to denote the average rate for coding the group using modes $\mathbf{v}$, and $\overline{D}_{\mathbf{v}}$ to denote the corresponding expected distortion, given the network conditions and the availability of packets for decoding. The Lagrangian cost associated with using a particular set of coding modes $\mathbf{v}$ is given by

$$J_{\mathbf{v}} = \overline{D}_{\mathbf{v}} + \lambda R_{\mathbf{v}}. \tag{6}$$

The optimal prediction modes for the group is determined by searching for the combination that results in minimal $J_{\mathbf{v}}$

$$\mathbf{v}_{\text{opt}} = \arg\min_{\mathbf{v}} J_{\mathbf{v}}. \tag{7}$$

In (6), λ is a Lagrange multiplier. We use the same λ as in Section III. At streaming time, when a prior frame that affects future dependency structure is negatively acknowledged by feedback, or time-out, the prediction dependency of the next frame to send is immediately changed by re-encoding using a most recently acknowledged frame as reference, or intra-coding. Error propagation is stopped in this way, and the prediction modes for the rest of the frames in the group are recomputed to avoid dependencies on any prior loss-afflicted frames.

Compared with the greedy approach, the prescient scheme may achieve global R-D optimality for a group of frames (which can be extended to the entire sequence). Within a particular group, as earlier frames are more important in the dependency structure than later frames, the prescient scheme applies stronger error-protection (by using higher modes) on earlier frames, compared to the equal protection given by the greedy scheme. Due to the weaker protection applied to later frames in a group, at lower costs, we expect bitrate savings by using the prescient approach.

B. Rate and Distortion Estimation for Generating R-D Preambles

In generating the R-D preambles and determining the optimal prediction modes using (6), accurate estimation of the rates and distortions is critical. The rates corresponding to different coding modes are obtained by encoding the frames with multiple trials offline. Denoting the rate of coding a particular frame i using mode v by $R_v[i]$, the average rate of the frames in a group with coding modes $\mathbf{v} = (v_0, v_1, \ldots, v_{L-1})$ is given by

$$R_{\mathbf{v}} = \frac{1}{L} \sum_{i=0}^{L-1} R_{v_i}[i]. \tag{8}$$

Estimation of the distortion is more challenging since the reconstruction distortion depends on the loss events of the packets. For a group of L packets, there exist 2^L loss patterns, and it is not feasible to measure and tabulate the distortion values for all combinations of packet losses. We estimate the distortion by using an accurate loss-distortion model we proposed in [48]. With this model, distortion values for general loss patterns can be extrapolated from a limited number of measurements.

This model explicitly considers the effect of different loss patterns, including burst losses and separated losses spaced apart by a lag, and accounts for the correlation between error frames. A simple loss concealment scheme is assumed where the lost frame is replaced by the previous frame at the decoder output. It is found in [48] that the distortion produced by a burst loss is generally greater than the sum of an equal number of single isolated losses, since it also includes cross-correlation terms between the error frames.

To avoid computing the distortion of all 2^L loss patterns, we approximate the overall expected distortion by only considering the distortion of single losses that are independent of each other, as well as the cross-term between any *two* losses. We ignore the distortion resulting from the interaction between more than two losses, since the probability of having more than two losses in a group is much smaller. In [49], it is shown that an approximation using first- and second-order Taylor expansion gives accurate estimation of the end-to-end distortion for packet loss rates of up to 20%. In this section, we base the distortion estimation on the model we established in [48]. For a particular group of L frames, the average mean-square distortion per frame is given by

$$\overline{D}_{\mathbf{v}} \approx \overline{D}_Q + \frac{1}{L}\left(\sum_{i=0}^{L-1} D_{\mathbf{v}}[i]\cdot p[i] + \sum_{i=0}^{L-2}\sum_{j=i+1}^{L-1} \Delta D_{\mathbf{v}}[i,j]\cdot p[i]\cdot p[j]\right) \tag{9}$$

where $\overline{D}_Q$ is the average distortion of the quantization error, and $p[i]$ is the loss probability of packet i. $D_{\mathbf{v}}[i]$ is the *total distortion* (not including the quantization error) of all the frames in the

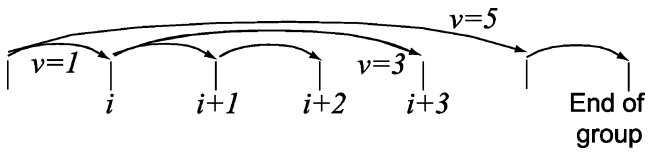


Fig. 4. Illustration of prediction modes.

group as a result of losing packet i, which includes not only the mean-square error (MSE) of Frame i, but also error propagation within the group. Due to the manipulation of packet dependency, $D_{\mathbf{v}}[i]$ is a function of the prediction modes $\mathbf{v}$. $D_{\mathbf{v}}[i]$ is computed by

$$D_{\mathbf{v}}[i] = \sigma^2[i] \cdot \alpha_{\mathbf{v}}[i] \tag{10}$$

where $\sigma^2[i]$ is the MSE of Frame i, and $\alpha_{\mathbf{v}}[i]$ is a factor that takes the error propagation into account. In the example illustrated in Fig. 4, $\alpha_{\mathbf{v}}[i] = 1 + r + r^2 + r$, where r $(r < 1)$ is a factor that accounts for the effect of spatial filtering in reducing the error power initially introduced. r mainly depends on the strength of the loop filter of the codec and is approximated by a constant average for a group or even for the entire sequence. In this example, the error introduced at Frame i is only propagated to three subsequent frames due to the prediction modes used.

In (9) $\Delta D_{\mathbf{v}}[i,j]$ is the cross-term in the total distortion as a result of losing both Frames i and j, which is not included in the distortions of single losses. $\Delta D_{\mathbf{v}}[i,j]$ is given by

$$\Delta D_{\mathbf{v}}[i,j] = 2\rho_{i,j} \cdot \sigma[i] \cdot \sigma[j] \cdot \alpha_{\mathbf{v}}[j] \tag{11}$$

where $\rho_{i,j}$ is the correlation coefficient between possible error Frames $j-1$ (propagated from Frame i), and j, more details of which can be found in [48]. If Frame $j-1$ is not afflicted by any error propagated from i, $\rho_{i,j} = 0$, due to the lack of any interaction between the two losses and the previous-frame-based concealment scheme.

To summarize, the three terms on the right-hand-side of (9) represent the quantization error free of frame losses, the distortion of single and independent losses, and the cross-term in the distortion of two losses, respectively. For a group of L frames, this approximation reduces the distortion computation from 2^L loss events down to $L + (1/2)L(L-1)$ events. The auxiliary information that needs to be premeasured and stored includes:

1) L MSE values $\sigma^2[i]$ for single losses at $i = 0, 1, \ldots, L-1$;
2) $(1/2)L(L-1)$ correlation coefficients $\rho_{i,j}$;
3) One attenuation factor r. The parameters in 1)–3) are to be used in (10), (11) and (9) to obtain $\overline{D}_{\mathbf{v}}$;
4) $(1/2)L(L+1)+1$ rate values, $R_v[i]$, for all eligible prediction modes, which are used in (8) to obtain $R_{\mathbf{v}}$.

C. Iterative Descent Algorithm to Determine the Optimal Coding Modes

With the coding structure introduced in Section IV-A, for a group of L frames, there exist $(L-1)!$ coding modes. The ideal way to determine the optimal $\mathbf{v}_{\text{opt}}$ in (7) is to loop over all the candidate coding modes and choose the combination that yields the minimal R-D cost in (6). While the complexity of this is intractably high, on the other hand, the interaction between the coding modes of different frames in a group prohibits optimizing each of them independently. To reduce the computational complexity, we use an *iterative descent algorithm* to find out the R-D optimized prediction modes for the group.

0 Initialize $\mathbf{v} = (0, 1, 1, ..., 1)$; compute $R_{\mathbf{v}}$ according to (8);
compute $\overline{D}_{\mathbf{v}}$ according to (9); $J_{\mathbf{v}}^{(0)} = \overline{D}_{\mathbf{v}} + \lambda R_{\mathbf{v}}$; $n = 1$.
1 Loop $i = 1 : L-1$
2 Loop $v_i = 0 : i$
3 Compute $R_{\mathbf{v}}$ according to (8);
compute $\overline{D}_{\mathbf{v}}$ according to (9); $J_{\mathbf{v}} = \overline{D}_{\mathbf{v}} + \lambda R_{\mathbf{v}}$.
4 End v_i
5 $v_{i,opt} = \arg\min_{v_i} J_{\mathbf{v}}$.
6 End i
7 $J_{\mathbf{v}}^{(n)} = \min J_{\mathbf{v}}$.
8 If $J_{\mathbf{v}}^{(n)} = J_{\mathbf{v}}^{(n-1)}$ stop; else $n = n+1$ and go to 1.
9 $\mathbf{v}_{opt} = (v_{0,opt}, v_{1,opt}, ..., v_{L-1,opt})$.

Fig. 5. Iterative descent algorithm to determine the optimal coding modes.

In this iterative approach, each time we minimize the Lagrangian cost $J_{\mathbf{v}}$ in (6) by varying one component in $\mathbf{v} = (v_0, v_1, \ldots, v_{L-1})$, while keeping all other components constant. We repeat this for every frame in the group and obtain the minimal cost in this iteration. We repeat the iteration until the cost does not further decrease. The complete algorithm is described in Fig. 5. In our simulations, this process for a group of 10 frames usually takes no more than four iterations to converge. Hence, the complexity reduces from $\mathcal{O}((L-1)!)$ to $\mathcal{O}(L)$.

V. Experimental Results

In this section we study the performance of the proposed network-adaptive scheme, when applied in the context of live encoding, and using prescient encoding for a group, respectively.

We compare the proposed scheme with a *P-I* scheme, in which normal P-frames are used with periodic I-frame insertions to combat packet loss. Note that the P-I scheme intrinsically provides certain amount of error resilience due to the use of periodic I-frames. We have implemented different schemes by modifying the H.264 TML 8.5. The video sequences used are *Foreman*, *Mother-Daughter*, *Salesman* and *Claire*; 230 frames are coded, and the frame rate is 30 fps. The forward and backward (feedback) channel have 1%

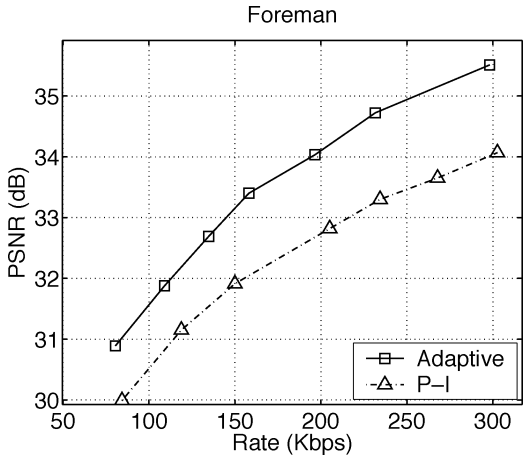


Fig. 6. R-D performance for *Foreman* sequence. $V = 5, \overline{p} = 10\%$.

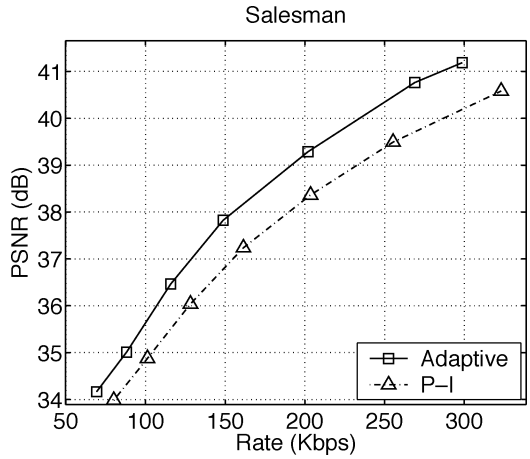


Fig. 8. R-D performance for *Salesman* sequence. $V = 5, \overline{p} = 10\%$.

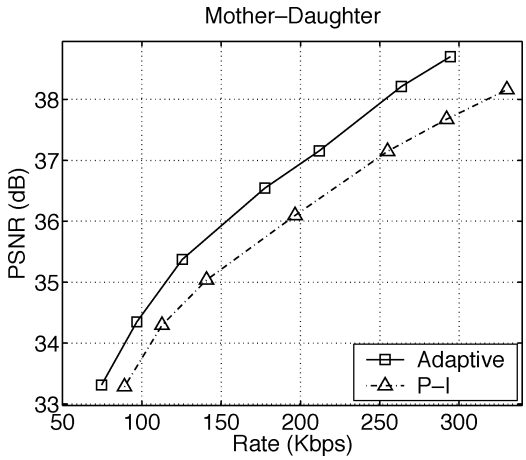


Fig. 7. R-D performance for *Mother-Daughter* sequence. $V = 5, \overline{p} = 10\%$.

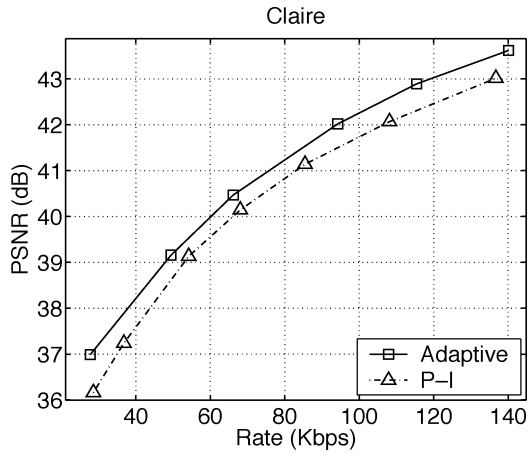


Fig. 9. R-D performance for *Claire* sequence. $V = 5, \overline{p} = 10\%$.

packet loss rate and a delay distribution density modeled using a shifted Gamma distribution, which is specified by a shift of 25 ms, a mean of 95 ms, and a standard deviation of 50 ms. The playout deadline is 165 ms. The network and streaming conditions above correspond to an effective overall loss rate of $\overline{p} = 10\%$, including the channel loss and the loss resulting from packets' late arrival. Coded frames are dropped according to the simulated network conditions and no retransmission is used due to the low-latency requirement. The PSNR of the decoded sequences is averaged over 30 random channel loss patterns. The first 30 frames are not included in the statistics to exclude the influence of the transient period.

A. Performance of Live Encoding

In live communication, Fig. 6 shows the R-D performance of transmitting the *Foreman* sequence over the network with the condition described above. The length of LTM is 5, and the distortion at different rates is obtained by varying the Q value and, hence, the Lagrange multiplier λ. The intra-rate in the P-I scheme is adjusted such that the rates keep up with the adaptive scheme at close Q values. A gain of 1.2 dB is observed at 200 kbps and 1.5 dB at 300 kbps by using the adaptive scheme, which corresponds to a bit rate saving of 35% at 34 dB. The gain is typically higher at higher rates since at lower rates LTM prediction with $v > 1$ is less efficient and the advantage of the adaptive scheme decreases. Fig. 7 shows the R-D performance of *Mother-Daughter* sequence under the same experimental conditions. A gain of 0.8 dB is observed at 200 kbps and 1.0 dB at nearly 300 kbps. The gain from using the adaptive scheme is lower compared to *Foreman* sequence since the effect of frame loss is smaller due to lower motion in the sequence. The results for *Salesman* and *Claire* sequences are also presented in Fig. 8 and 9 respectively. The lowest gain is found in *Claire* sequence due to the near-static nature of the sequence.

Using different sets of parameters for the Gamma delay distribution, distortion at different effective channel loss rates is obtained and shown in Fig. 10 for *Foreman* encoded at approximately the same bitrate of 200 kbps using the two schemes. The gain is observed ranging from 0.7 to 1.8 dB, depending on the channel loss rate. The advantage of using the adaptive scheme is more obvious at higher channel loss rate. The R-D performance with different LTM length V is shown in Fig. 11. At a feedback delay of seven frames, an increase of V from 2 to 5 results in 0.5–0.7-dB gain at higher rates while an increase from 7 to 8 does not give much further improvement. This gives us some idea on how to choose the LTM length for the trade-off between performance and storage complexity.

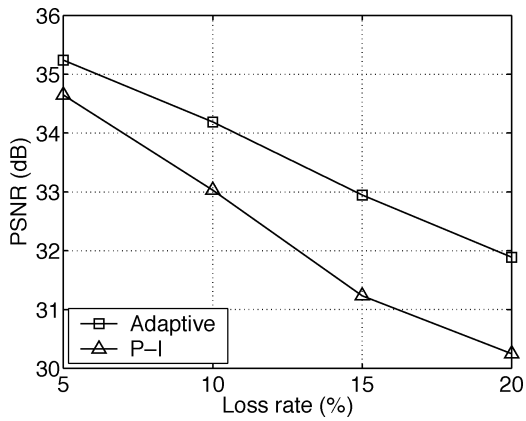


Fig. 10. Distortion at different channel loss rates. *Foreman* sequence. $V = 5$.

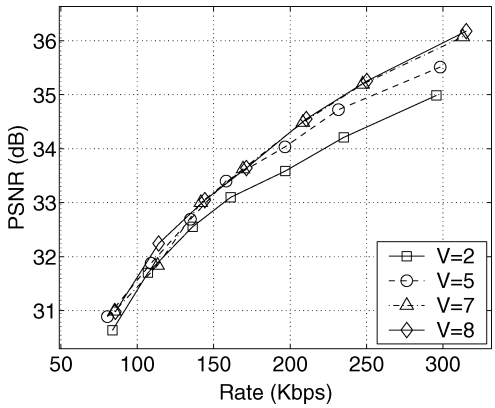


Fig. 11. R-D performance at different LTM lengths. *Foreman* sequence. $\overline{p} = 10\%$.

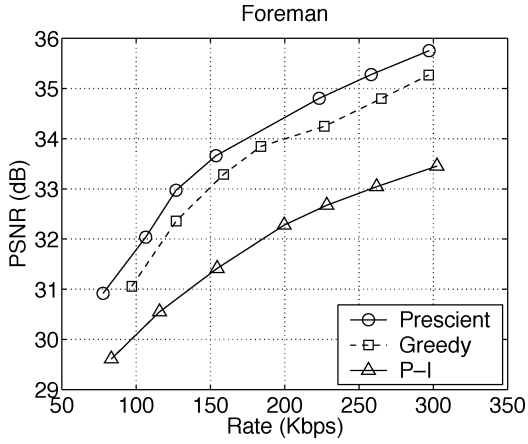


Fig. 12. R-D performance for *Foreman* sequence. $L = 10$.

B. Performance of the Prescient Scheme

We compare the performance of the proposed scheme (*prescient*) with the P-I scheme, as well as the *greedy* scheme shown in Subsection V-A. For the prescient scheme, the group size is $L = 10$ frames and for the greedy scheme $V = 10$.

Fig. 12 shows the R-D performance for *Foreman* sequence. Comparing the prescient scheme and the P-I scheme, a gain of 1.7 dB is observed at 200 kbps and 1.8 dB at 300 kbps, which corresponds to a bit rate saving of 42% at 33 dB. Fig. 13 shows

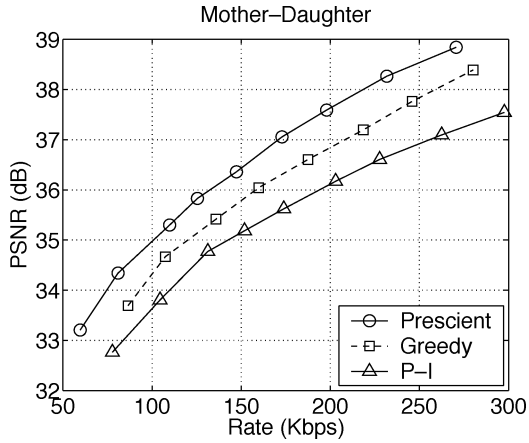


Fig. 13. R-D performance for *Mother-Daughter* sequence. $L = 10$.

the results for *Mother-Daughter* under the same experimental conditions. A gain of 0.7 dB is observed at 200 kbps.

Comparing the prescient scheme with our previous greedy scheme, we observe typical gains of 0.5 dB for *Foreman* and 0.6 dB for *Mother-Daughter*. This corresponds to a bit rate saving of 16% at 31 dB for *Foreman*, and 20% at 34 dB for *Mother-Daughter*. The savings in bitrate, especially at lower rates, can be explained by the weaker error-resilience applied for later frames in a group and the corresponding lower rate cost. The complexity of the prescient scheme is also much lower for the streaming of prestored video.

VI. Conclusion

In this paper, error-resilient source coding techniques are proposed for real-time and on-demand video communications that impose very stringent delay requirements. A novel scheme of dynamic management of the dependency across packets is proposed using optimal RPS at the frame level that adapts to the network. The optimal selection of the reference picture is achieved within a rate-distortion framework, which minimizes the expected end-to-end distortion given a rate constraint. Experiments demonstrate that the proposed scheme provides significant performance gains over a simple intra-insertion scheme, with typical gains of 0.5–1.5 dB.

For streaming of pre-encoded media, we may take advantage of optimizing the dependency of a group packets to send in the future. Compared with the greedy scheme for live-streaming, in which case optimality is only achieved locally, the prescient optimization scheme achieves global optimality for a group of packets, as well as improved R-D performance. Experiments demonstrate that the prescient scheme provides consistent bitrate savings of typically 15%–20% compared to the greedy scheme.

Using network-adaptive coding, the dependency across media packets is manipulated and optimized, and the need for retransmitting lost packets is eliminated. This enables VoIP-like below one second low-latency for video communication, with good video quality maintained. Reduced latency significantly improves user interactivity for conversational and on-demand video applications.

ACKNOWLEDGMENT

The authors would like to thank M. Flierl and R. Zhang for helpful discussions.

REFERENCES

[1] N. Johnson. (1998) The Case for Standard Real-Time Video, a GTS White Paper. [Online]. Available: http://www.downrecs.com/pdf/gts.pdf

[2] Y. J. Liang, N. Färber, and B. Girod, "Adaptive playout scheduling and loss concealment for voice communication over IP networks," *IEEE Trans. Multimedia*, vol. 5, no. 4, pp. 532–543, Dec. 2003.

[3] S. Wenger, G. D. Knorr, J. Ott, and F. Kossentini, "Error resilience support in H.263+," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 7, pp. 867–877, Nov. 1998.

[4] R. Talluri, "Error-resilient video coding in the ISO MPEG-4 standard," *IEEE Commun. Mag.*, vol. 36, no. 6, pp. 112–119, Jun. 1998.

[5] N. Färber, B. Girod, and J. Villasenor, "Extension of ITU-T Recommendation H.324 for error-resilient video transmission," *IEEE Commun. Mag.*, pp. 120–128, Jun. 1998.

[6] B. Girod and N. Färber, "Feedback-based error control for mobile video transmission," *Proc. IEEE*, vol. 87, no. 10, pp. 1707–1723, Oct. 1999.

[7] "Special Issue on streaming video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, Mar. 2001.

[8] "Special Issue on error resilient image and video transmission," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 6, Jun. 2001.

[9] G. J. Conklin, G. S. Greenbaum, K. O. Lillevold, A. F. Lippman, and Y. A. Reznik, "Video coding for streaming media delivery on the Internet," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 269–281, Mar. 2001.

[10] B. Girod, M. Kalman, Y. J. Liang, and R. Zhang, "Advances in channel-adaptive video streaming," *Wireless Commun. Mobile Comput.*, vol. 2, no. 6, pp. 573–584, Sep. 2002.

[11] G.J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Process. Mag.*, vol. 15, no. 6, pp. 74–90, Nov. 1998.

[12] A. Ortega and K. Ramchandran, "From rate-distortion theory to commercial image and video compression technology," *IEEE Signal Process. Mag.*, vol. 15, no. 6, pp. 20–22, Nov. 1998.

[13] T. Wiegand, X. Zhang, and B. Girod, "Long-term memory motion-compensated prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 1, pp. 70–84, Feb. 1999.

[14] T. Wiegand, N. Färber, and B. Girod, "Error-resilient video transmission using long-term memory motion-compensated prediction," *IEEE J. Sel. Areas. Commun.*, vol. 18, no. 6, pp. 1050–1062, Jun. 2000.

[15] P. A. Chou, A. E. Mohr, A. Wang, and S. Mehrotra, "Error control for receiver-driven layered multicast of audio and video," *IEEE Trans. Multimedia*, vol. 3, no. 1, pp. 108–122, Mar. 2001.

[16] S. B. Wicker, *Error Control Systems for Digital Communication and Storage*. Englewood Cliffs, NJ: Prentice Hall, 1995.

[17] M. Khansari, A. Jalali, E. Dubois, and P. Mermelstein, "Low bit-rate video transmission over fading channels for wireless microcellular systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 1, pp. 1–11, Feb. 1996.

[18] B. Dempsey, J. Liebeherr, and A. Weaver, "On retransmission-based error control for continuous media traffic in packet-switching networks," *Comput. Netw. ISDN Syst. J.*, vol. 28, no. 5, pp. 719–736, Mar. 1996.

[19] C. Papadopoulos and G. M. Parulkar, "Retransmission-based error control for continuous media applications," presented at the *Proc. Network and Operating System Support for Digital Audio and Video (NOSSDAV)*, Zushi, Japan, Jul. 1996.

[20] H. Liu and M. El Zarki, "Performance of H.263 video transmission over wireless channels using hybrid ARQ," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 9, pp. 1775–1786, Dec. 1999.

[21] J. Y. Liao and J. D. Villasenor, "Adaptive intra update for video coding over noisy channels," in *Proc. IEEE Int. Conf. Image Processing*, vol. 3, Lausanne, Switzerland, Sep. 1996, pp. 763–766.

[22] R. O. Hinds, T. N. Pappas, and J. S. Lim, "Joint block-based video source/channel coding for packet-switched networks," in *Proc. SPIE VCIP*, vol. 3309, San Jose, CA, Oct. 1998, pp. 124–133.

[23] G. Cote and F. Kossentin, "Optimal intra coding of blocks for robust video communication over the Internet," *Signal Process. Image Commun.*, vol. 15, no. 1–2, pp. 25–34, Sep. 1999.

[24] R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE J. Sel. Areas. Commun.*, vol. 18, no. 6, pp. 966–976, Jun. 2000.

[25] "RTP Payload Format for MPEGI/MPEG2 Video," Internet Engineering Task Force, RFC-2250, Jan. 1998.

[26] A. Albanese, J. Blömer, J. Edmonds, M. Luby, and M. Sudan, "Priority encoding transmission," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 1737–1744, Nov. 1996.

[27] P. C. Cosman, J. K. Rogers, P. G. Sherwood, and K. Zeger, "Image transmission over channels with bit errors and packet erasures," in *Proc. 32nd Asilomar Conf. Signals, Systems and Computers*, vol. 2, Pacific Grove, CA, Nov. 1998, pp. 1621–1625.

[28] W. Tan and A. Zakhor, "Video multicast using layered FEC and scalable compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 373–387, Mar. 2001.

[29] P. A. Chou and Z. Miao, "Rate-distortion optimized streaming of packetized media," *IEEE Trans. Multimedia*, submitted for publication.

[30] S. Fukunaga, T. Nakai, and H. Inoue, "Error resilient video coding by dynamic replacing of reference pictures," in *Proc. IEEE Global Telecommunications Conf.*, vol. 3, London, U.K., Nov. 1996, pp. 1503–1508.

[31] *Video Coding for Low Bitrate Communication*, ITU-T Recommendation 1-1.263 Version 2 (H.263+), Jan. 1998.

[32] *Advanced Video Coding (AVC) for Generic Audiovisual Services*, ITU-T Recommendation H.264, May 2003.

[33] *Information Technology—Coding of Audio-Visual Objects: Visual (MPEG-4)*, ISO/IEC JTC1/SC29/WG11 Final Committee Draft 14496-2, Mar. 1998.

[34] M. Budagavi and J. D. Gibson, "Multiframe video coding for improved performance over wireless channels," *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 252–265, Feb. 2001.

[35] Z. Miao and A. Ortega, "Optimal scheduling for streaming of scalable media," in *Proc. 34th Asilomar Conf. Signals, Systems and Computers*, vol. 2, Pacific Grove, CA, Nov. 2000, pp. 1357–1362.

[36] J. Chakareski, P. A. Chou, and B. Aazhang, "Computing rate-distortion optimized policies for streaming media to wireless clients," in *Proc. Data Compression Conf.*, Snowbird, UT, Apr. 2002, pp. 53–62.

[37] A. Sehgal and P. A. Chou, "Cost-distortion optimized streaming media over DiffServ networks," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME)*, vol. 1, Lausanne, Switzerland, Aug. 2002, pp. 857–860.

[38] M. Kalman, E. Steinbach, and B. Girod, "R-D optimized media streaming enhanced with adaptive media playout," in *Proc. IEEE Int. Conf. Multimedia*, vol. 1, Lausanne, Switzerland, Aug. 2002, pp. 869–872.

[39] M. Kalman, P. Ramanathan, and B. Girod, "Rate-distortion optimized streaming with multiple deadlines," in *Proc. IEEE Int. Conf. Image Processing*, vol. 3, Barcelona, Spain, Sep. 2003, pp. 661–664.

[40] B. Girod and N. Färber, "Feedback-based error control for mobile video transmission," *Proc. IEEE*, vol. 87, no. 10, pp. 1707–1723, Oct. 1999.

[41] S. Lin, S. Mao, Y. Wang, and S. Panwar, "A reference picture selection scheme for video transmission over ad hoc networks using multiple paths," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME)*, Aug. 2001, pp. 96–99.

[42] Y. J. Liang, M. Flierl, and B. Girod, "Low-latency video transmission over lossy packet networks using rate-distortion optimized reference picture selection," in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, Rochester, NY, Sep. 2002, pp. 181–184.

[43] H.26L Test Model Long Term Number 8, ITU-T Video Coding Expert Group. (2001, Jul.). [Online]. Available: ftp://standard.pictel.com/video-site/h26L/tml8.doc

[44] T. Wiegand and B. Girod, "Lagrange multiplier selection in hybrid video coder control," in *Proc. IEEE Int. Conf. Image Processing*, vol. 3, Thessaloniki, Greece, Oct. 2001, pp. 542–545.

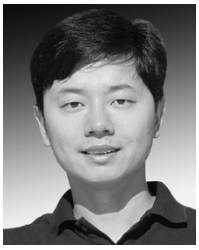
[45] R. Zhang, S. L. Regunathan, and K. Rose, "Prescient mode selection for robust video coding," in *Proc. IEEE Int. Conf. Image Processing*, vol. 1, Thessaloniki, Greece, Oct. 2001, pp. 974–977.

[46] Y. J. Liang and B. Girod, "Prescient R-D optimized packet dependency management for low-latency video streaming," in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, Barcelona, Spain, Sep. 2003, pp. 659–662.

[47] ——, "Low-latency streaming of pre-encoded video using channel-adaptive bitstream assembly," in *Proc. IEEE Int. Conf. Multimedia and Expo* , vol. 1, Lausanne, Switzerland, Aug. 2002, pp. 873–876.

[48] Y. J. Liang, J. G. Apostolopoulos, and B. Girod, "Analysis of packet loss for compressed video: Does burst-length matter?," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 5, Hong Kong, Apr. 2003, pp. 684–687.

[49] R. Zhang, S. L. Regunathan, and K. Rose, "End-to-end distortion estimation for RD-based robust delivery of pre-compressed video," in *Proc. 35th Asilomar Conf. Signals, Systems and Computers*, vol. 1, Pacific Grove, CA, Nov. 2001, pp. 210–214.



Yi J. Liang (M'02) received the B.Eng. degree from Tsinghua University, Beijing, China, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 2003.

He is currently holding positions at Qualcomm CDMA Technologies, San Diego, CA, and is responsible for video and multimedia system design and development for Qualcomm's mobile station modem (MSM) chipsets. His expertise is in the areas of networked multimedia systems, real-time voice and video communication, and low-latency media streaming over the wire-line and wireless networks. From 2000 to 2001, he conducted research with Netergy Networks, Inc., Santa Clara, CA, on voice over IP systems that provide improved quality over best-effort networks. From 2001 to 2003, he had been the lead of the Stanford—Hewlett-Packard Laboratories low-latency video streaming project, in which he and his colleagues developed error-resilience techniques for rich media communication over IP networks at low latency. In the summer of 2002 at Hewlett-Packard Laboratories, Palo Alto, CA, he developed an accurate loss-distortion model for compressed video and contributed in the development of the mobile streaming media content delivery network (MSM-CDN) that delivers rich media over 3G wireless.

Bernd Girod (S'80–M'80–SM'97–F'98) received the M. S. degree in electrical engineering from Georgia Institute of Technology, Atlanta, in 1980 and the Ph.D. degree (with highest honors) from the University of Hannover, Hannover, Germany, in 1987.

He is presently a Professor of Electrical Engineering with the Information Systems Laboratory, Stanford University, Stanford, CA. He also holds a courtesy appointment with the Stanford Department of Computer Science and he serves as Director of the Image Systems Engineering Program at Stanford. His research interests include networked media systems, video signal compression and coding, and three-dimensional image analysis and synthesis. Until 1987, he was a Member of the Research Staff with the Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung, University of Hannover, working on moving image coding, human visual perception, and information theory. In 1988, he joined Massachusetts Institute of Technology, Cambridge, first as a Visiting Scientist with the Research Laboratory of Electronics and then as an Assistant Professor of Media Technology at the Media Laboratory. From 1990 to 1993, he was a Professor of Computer Graphics and Technical Director of the Academy of Media Arts, Cologne, Germany, jointly appointed with the Computer Science Section of Cologne University. He was a Visiting Adjunct Professor with the Digital Signal Processing Group, Georgia Institute of Technology, Atlanta, in 1993. From 1993 until 1999, he was the Chaired Professor of Electrical Engineering/Telecommunications, University of Erlangen-Nuremberg, Nuremberg, Germany, and the Head of the Telecommunications Institute I, codirecting the Telecommunications Laboratory. He has served as the Chairman of the Electrical Engineering Department from 1995 to 1997 and as Director of the Center of Excellence "3-D Image Analysis and Synthesis" from 1995 to 1999. He was a Visiting Professor with the Information Systems Laboratory of Stanford University during the 1997–1998 academic year. As an entrepreneur, he has worked successfully with several start-up ventures as founder, investor, director, or advisor. Most notably, he has been a cofounder and Chief Scientist of Vivo Software, Inc., Waltham, MA (1993–1998); after Vivo's aquisition, Chief Scientist of RealNetworks, Inc. (1998–2002), and an outside Director of 8×8, Inc. (1996–2004). He has authored or coauthored one major textbook, two monographs, and over 250 book chapters, journal articles, and conference papers in his field, and he holds about 20 international patents.

Prof. Girod has been a member of the IEEE Image and Multidimensional Signal Processing Committee from 1989 to 1997. He was named "Distinguished Lecturer" in 2002 by the IEEE Signal Processing Society. Together with J. Eggers, he was the recipient of the 2002 EURASIP Best Paper Award. He has served on the Editorial Boards or as an Associate Editor for several journals in his field and is currently Area Editor for Speech, Image, Video Signal Processing of the IEEE TRANSACTIONS ON COMMUNICATIONS. He has served on numerous conference committees, e.g., as Tutorial Chair of ICASSP-97 in Munich, Germany, and ICIP-2000 in Vancouver, ON, Canada, as General Chair of the 1998 IEEE Image and Multidimensional Signal Processing Workshop in Alpbach, Austria, and as General Chair of the Visual Communication and Image Processing Conference (VCIP) in San Jose, CA, in 2001.