Minimizing the Quality-of-Service Requirement for Real-Time Video Conferencing*

(Extended abstract)

Injong Rhee, Sarah Chodrow, Radhika Rammohan,
Shun Yan Cheung, and Vaidy Sunderam

Department of Mathematics and Computer Science
Emory University
Atlanta, GA 30322

November 1996

Abstract

This paper explores video transmission over networks that allow the reservation of guaranteed bandwidth, such as ATM[7] and RSVP[9]. Because the reservation of bandwidth is costly, we would like to minimize the amount of reserved bandwidth while utilizing the available best-effort bandwidth as much as possible. However, the coding scheme has to compensate for the expected data loss from the best-effort channel to avoid image degradation. Some earlier schemes (H.261, MPEG) require high reserved bandwidth to maintain good image quality, because they do not compensate for lost data. Other schemes (Motion JPEG, intra-H.261) do compensate for data loss. However, they tend to increase the total data rate. We present a reservation-based coding (RVC) which is a variant of the commonly-used video conferencing standard H.261[4]. RVC compensates for data loss in the best-effort channel, without overly increasing the total data rate. This minimizes the reserved bandwidth needed to maintain a high quality video conference. Our experimental results show that the bandwidth required for the reserved channel is minimal (averaging 1K-2K bytes per medium to high motion CIF frame) while maintaining good image quality under data loss. Further, the total bandwidth requirements for an entire frame is only slightly higher than that of H.261, and much less than that of intra-H.261. The RVC coding scheme shows a good tradeoff between data rate and tolerance of data loss. RVC's overall data rate is only slightly higher than H.261, and it exhibits excellent tolerance to data loss. Therefore, under the RVC scheme, the amount of reserved bandwidth can be minimized. Further, RVC's total data rate is up to 20% less than that of intra-H.261, while maintaining comparable image quality under data loss.

1 Introduction

New networking paradigms such ATM [7] and RSVP [9] provide support for performance guarantees needed by real-time applications, such as video conferencing, to ensure a minimum level of performance. This is achieved by reserving network resources (e.g., buffers and/or bandwidth) on the communication paths. Reserved resources are not shared among different connections and it is therefore essential to keep reservation at a minimum. In order to minimize reservation, video transmission schemes will typically send essential video information on the reserved channel and add-on signals with best effort transmission. Existing coding and transmission methods either suffer from error propagation due to inter-frame dependencies or require an excessive amount of bandwidth by encoding frames independently. In this paper, we present RVC, a novel reservation-based video coding and transmission method to split a video stream into an essential and

*Research supported in part by NSF grant ASC-9527186

an add-on component that reduces inter-frame dependencies in the video sequence while keeping needed reservation to a minimum.

Video signals contain a high degree of variance which necessitates the sending of a large amount of data over the network. Video conferencing applications can tolerate a certain amount of signal loss, albeit with some degradation of image quality. However, video conferencing applications cannot tolerate delays in the transmission of the video signal. By reserving some guaranteed bandwidth, a video application can meet these real-time constraints.

Guaranteed bandwidth is expensive and it is important to identify and reserve only the minimal amount of bandwidth needed. Compression can significantly reduce the minimum bandwidth required. Furthermore, a hierarchical compression method can compress a video signal into separate layers of signals [6] that can be combined at the receiver. Typically, the hierarchical coder outputs essential, or base streams and non-essential streams that refine the coded signal in the essential streams incrementally. The advantage of partitioning video into essential and non-essential signals, is is that reserved bandwidth can be kept to a minimum by sending the refinement signals via an available best-effort channel.

Many different types of hierarchical coding exists. Two popular methods partition the image data over the temporal and spatial domains, respectively. In a temporally hierarchically encoded transmission [2], a periodic subsequence of frames is sent over the reservation channel, and the remaining image frames are sent using best-effort. In a typical scenario, the reservation channel carries I-frames (a full refresh image frame) while the best-effort channel carries predictive frames (motion and error correction data). Spatial domain coders [3, 8, 1] divide their output divided into a meaningful low-resolution image and additional enhancement information. The reservation channel carries the *base stream* containing the low-resolution image–e.g., MPEG headers and the first few DCT coefficients. Both temporal and spatial domain coding suffer from the problem of error propagation due to data loss, since the decoding of some frames depends on the availability of a preceding frame. Since the video signal is split between reservation and best-effort channels, some preceding frame (or portions thereof) may not be available when the receiver decodes the current frame, thus introducing errors. These errors propagate because all later frames in the subsequence use the erroneous frame as a reference in their decoding.

H.261 [4] is a video conferencing standard currently in common use. A major drawback of H.261 is that the fact that the decoding of P- and B-frames relies on the availability of *all* information in a preceding frame. As a result, decoding errors due to loss of partial data in an earlier frame in a sequence will *propagate* through all later frames in that sequence. Other H.261-based hierarchical encoding schemes display high degradation of image quality because of their sensitivity to data loss. To minimize data loss to ensure good quality of video, full bandwidth reservation would be necessary in H.261. The new RVC video coding scheme proposed here is based on H.261 and uses spatial partitioning to encode video into an essential and other non-essential streams. The novelty of the RVC approach is the fact that the decoding of information in the essential stream relies only on the availability of the *essential* information in a preceding frame. To ensure good quality video with the RVC method, bandwidth need only be reserved for the base stream. Non-essential data from a frame that are lost will result in decoding errors *only* within that frame and will not affect any other frame. Experimental results show that video encoded in RVC is much less sensitive to data loss than other H.261-based methods. Furthermore, the added degree of robustness was achieved without incurring much much bandwidth overhead.

Section 2 contains a detailed description of our scheme. In Section 3, we present the experimental results, and conclude in Section 4.

2 Video Coding

Below, we briefly discuss CCITT recommendation H.261 [4] the video conferencing standard upon which many internet-based video conferencing schemes are built, then explain the pitfalls of existing coding schemes in the event of packet loss, and describe our reservation based coding scheme in detail.

H.261 (see Figure 1) describes video encoding and decoding methods for video stream transmission at rates of $p \times 64$ kbits/s, where $1 \leq p \leq 30$. The first video frame (an I-frame) is divided into fixed-size blocks, that are then transformed by discrete cosine transform (DCT) into a block of orthogonal frequency coefficients. These coefficients are quantized (Q), then fed into a variable-length coder (ZZ/VLC) and

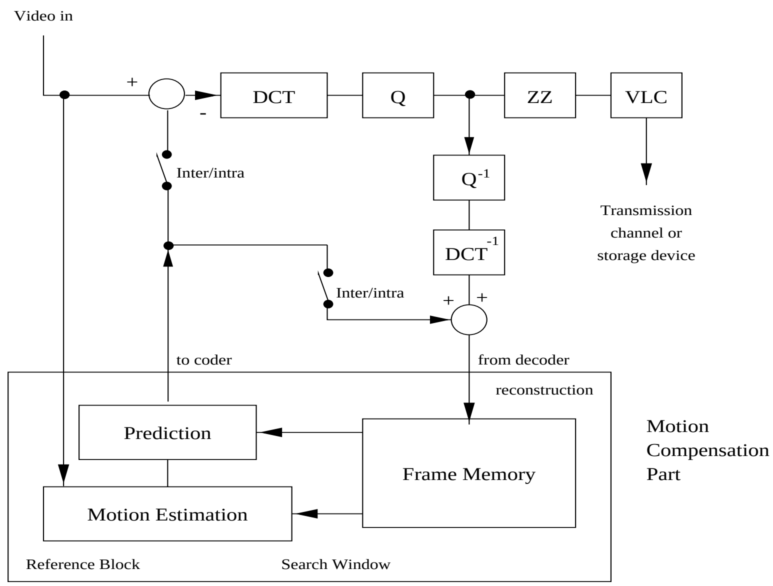


Figure 1: The H.261 video encoding scheme.

transmitted over the network. The frame is also reconstructed as it would be decoded by the receiver (Q^{-1}, DCT^{-1}). The following frame (a predictive, or P-frame) is compared to the previous *reconstructed* frame to generate a motion vector for each block. A new image is constructed solely from the motion vector of every block. The difference (compensation error) between the current frame and the new image is encoded in the same manner as the I-frame. The motion vector and encoded compensation error for each block are then transmitted. I-frames are also periodically sent to refresh the image. Note that the motion vector and compensation error for each frame are generated from the entire previous frame.

New network paradigms, such as ATM and RSVP, allow an application to reserve a certain amount of network bandwidth. The reserved bandwidth can be used to send a time critical data such as real-time video or audio signals. However, because of the high data rate of real-time video, it is impractical to reserve bandwidth for the entire video signal. Instead, one can divide the video signal hierarchically, into essential data and incremental refinements, and reserve bandwidth only for the essential data (e.g., VBR in ATM). The refinements can be transmitted over an available best-effort channel (e.g., ABR/UBR in ATM). This way, video images can be decoded and displayed in a timely fashion using only the essential data while the refinements are used to enhance the quality of the decoded image (if they are received in time).

Temporal domain encoding schemes [2] periodically send I-frames over the reserved channel, and send the remaining P-frames over the best-effort channel. In internet-based video conferencing, one cannot transmit I-frames frequently because of the lack of total available bandwidth. Because of the long time span between two consecutive I-frames sent over the reserved channel (30 to 100 frames), if one of P-frames contains some error due to lost packets, all the P-frames that depends on the lost frame contain the same error. Thus, the image quality degrades rapidly, because the loss of one P-frame propagates error until the next I-frame is sent.

Spatial domain encoding schemes [8] divide the DCT coefficients for each block into low-frequency and high-frequency groups. The low frequency coefficients and the motion vector are sent over the reserved channel; the remaining coefficients are transmitted across the best-available channel. This can be easily implemented on top of H.261 or MPEG. However, if implemented without care, the scheme suffers from the same image degradation as in TDC. This is because in H.261 each P-frame is coded based on the entire information about its previous frame. Motion compensation errors of a P-frame which is encoded through DCT are derived from the difference between the decoded *lossless* image of the previous frame and the image currently being encoded. Thus, if any information about the previous frame is lost, its next frame carries the error due to the loss. Again, these errors propagate until the image is refreshed by the next I-frame. In this scheme, in order to sustain a good quality of video images, more information has to be transmitted over the reserved channel, requiring to reserve a larger bandwidth.

A third scheme (Intra-H.261) replaces motion estimation with conditional replenishment [5]. Each block is compared to the corresponding block in the previous frame. If the blocks are sufficiently different, then the block is encoded. Otherwise, the block is not encoded. Since each encoded block is independent of previous blocks, errors in a frame are confined to that frame and do not propagate very much. However, the data rate is high because it does not fully take advantage of the temporal redundancy present in most video signals.

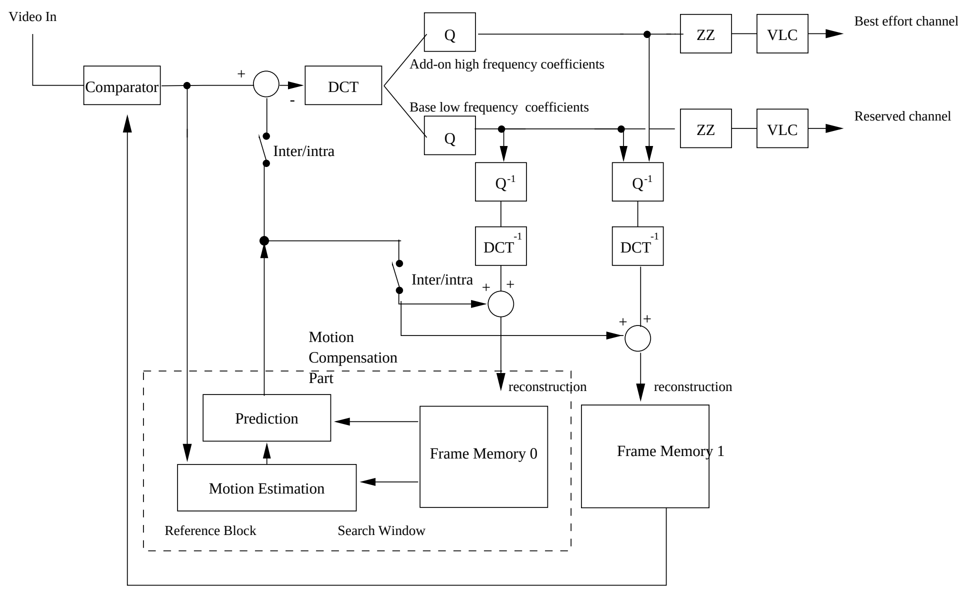


Figure 2: Reservation-based video coding scheme.

Our scheme, called Reservation-based coding (RVC), takes advantage of motion estimation to reduce the data rate, while minimizing error propagation due to packet loss. Figure 2 illustrates our scheme. The encoder retains two reference frames. The first frame (Frame Memory 0) is reconstructed only from information sent on the reserved channel, and it is this frame that is used to compute the motion vector and compensation error. The second frame (Frame Memory 1) is reconstructed from the previous frame's compensation error and motion vector, and is compared to the current frame to perform conditional replenishment. The sender and receiver have the same image history, because all computation is done only on reliably transmitted data. As as result, errors due to packet loss in the best effort channel are restricted to a single frame and do not propagate.

The computational costs and bandwidth requirements of RVC are comparable to previous schemes. Generating the second frame merely copies the results of the inverse DCT and does not require additional expensive computation. Computation can be further reduced by using the previous frame for conditional replenishment–this alleviates the need for the second frame, but theoretically can introduce propagated errors. The decoding error that occurs in abrupt-motion regions may not be visible at the time of conditional replenishment when using the original rather than the decoded frame. However, motion in video conferencing usually continues in the same region through later frames; therefore the motion present in the later frames will replenish that block. This optimization will reduce the computational costs of encoding below those of H.261, because only low-frequency coefficients are decoded in the encoder.

One concern raised by RVC is that the reference frame in Frame Memory 0 is constructed from incomplete information of the previous image. Thus, motion prediction error is theoretically larger in RVC than other schemes, resulting in an increased data rate. However, experimental results show that the reference frame does contain much of the image content of previous frames, and combined with the benefits of motion compensation, results in a data rate comparable to H.261.

3 Experimental Results

In this section, we present a series of experiments to evaluate RVC and to compare it to other video conferencing schemes. We compare RVC to temporal domain coding (TDC), and to spatial domain coding using H.261 as in IVS [8], and intra-H.261 as in vic [5].

3.1 Experimental Environment

We have implemented all four schemes by modifying IVS[1] Motion estimation under IVS is very rudimentary. The comparator checks for sufficient motion between the current block and the previous block in the same position. If the motion threshold has been passed, then the difference between those blocks is encoded, but no motion vector is sent.

RVC, intra-H.261 and H.261 divide each coded block into a base stream and an optional stream. In our implementations, the base stream consisted of the DC value and the first two DCT coefficients, and the optional stream contained the remaining 61 coefficients. TDC divides a video sequence temporally into I-frames and P-frames, with an I-frame transmitted every 50 frames, and when the image changes sufficiently.

We created two 500 frame video sequences in CIF format (352×288 pixels)–one with a high degree of motion, the other with a low degree of motion. Encoded images were transmitted over a local area network, with the decoder and encoder running on UltraSparc processors. The two sequences were coded by all four schemes under various conditions.

We dropped packets randomly from the best-effort channel, based on a pre-specified loss rate. Because our experiments ran on a local area network between fast processors, we neither expected nor observed much packet loss due to the network itself; rather, all observed behavior resulted from the controlled environment of the experiment. There was no packet loss observed on the reliable channel.

3.2 Size of the reserved channel

We ran our codecs on the two video sequences to obtain statistics for the bandwidth (in bytes/frame) to be reserved for each scheme. We observed the following behaviors: The average size of the base stream under the spatial domain coding schemes (RVC, intra-H.261, H.261) was 1.2 kbytes/frame for a low motion sequence, and 2.7 kbytes/frame for a high motion sequence. TDC using one I-frame every 50 frames generated 15-20 kbytes/I-frame. Because TDC does not use motion estimation, there is no differentiation between the high and low motion sequences.

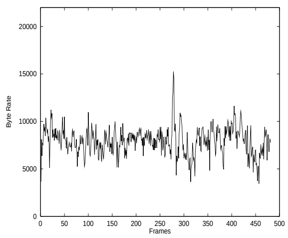
3.3 Total data rate with conditional replenishment

Because RVC has the potential to increase the overall data rate (due to the lower fidelity of the base stream), we are interested in measuring the increase, if it arises. Indeed, we observed that RVC does not increase the total data rate much, and exhibits a data rate considerably lower than intra-H.261.

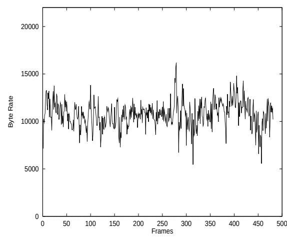
In this experiment, we compared RVC with intra-H.261, H.261, and TDC. Figure 3 shows the relative *byte*-rates of the four schemes for the low-motion sequence of 500 frames. As expected, intra-H.261 has the highest data rate because each block is intra-coded without taking advantage of temporal redundancy. RVC shows a higher data rate than H.261 because the reference frame used in motion estimation only contains partial information about the previous image, namely its base stream. However, the increase is small. H.261 and TDC exhibit similar behavior because they send I-frames at about the same rate (note the peaks every 50 frames under TDC).

Figure 4 shows the relative byte-rates of the four schemes for the high-motion sequence of 500 frames. The rudimentary motion estimation implemented in IVS is ineffective for high-motion sequences. The slightly better behavior of RVC in this instance is probably due to an artifact of the data. With better motion estimation, we expect results similar to the low-motion sequence.

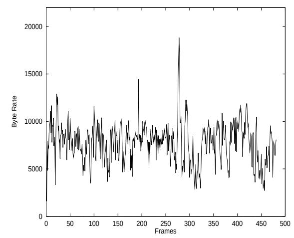
[1] source available from `http://www.inria.fr/rodeo/personnel/Thierry.Turletti/ivs.html`.



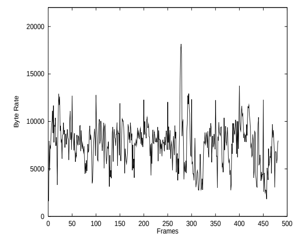
(a) RVC (avg = 8041)



(b) intra-H.261 (avg = 10855)

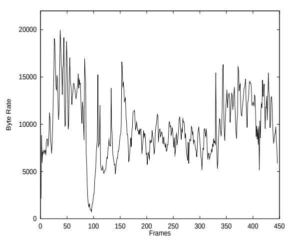


(c) H.261 (avg = 7853)

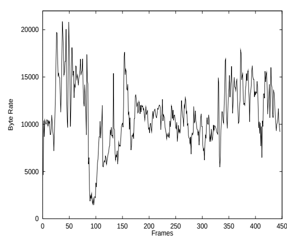


(d) TDC (avg = 7743)

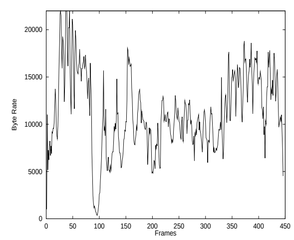
Figure 3: Byte rate/frame for low-motion sequence.



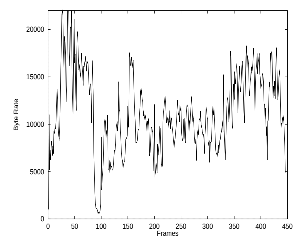
(a) RVC (avg = 9800)



(b) intra-H.261 (avg = 11135)



(c) H.261 (avg = 11363)



(d) TDC (avg = 11343)

Figure 4: Byte rate/frame for high-motion sequence.

3.4 Image quality under various data loss rates

We believe that RVC's image quality is comparable to that of intra-H.261, while incurring a lower data rate, and that RVC has higher image quality under data loss than H.261 and TDC.

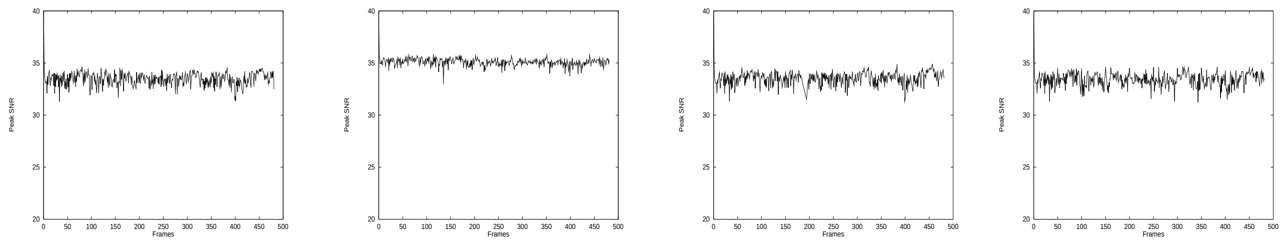
We compared the four coding schemes under conditions of no data loss, 10% loss, 20% loss, and 30% loss on both the low motion and high motion sequences.

Figure 5 shows the peak signal-to-noise ratio for each scheme across the low motion sequence. Intra-H.261 has the highest average image quality throughout, because each image frame is coded almost independently of previous frames. In general, intra-coding produces better image quality than motion-based coding. Under increasing data loss, however, the image quality of RVC approaches that of intra-H.261. TDC does not tolerate data loss well. As packets in the P-frames are lost, the image degrades until the next I-frame is received. (Note the peaks every 50 frames.) H.261's image quality degrades faster than RVC as the data loss increases.

Figure 6 shows the peak signal-to-noise ratio for each scheme across the high motion sequence. As the data loss rate increases, H.261 degrades even more quickly than in the low-motion sequence. RVC and intra-H.261 exhibit behavior similar to the low-motion sequence. Image quality is marginally worse. This indicates that RVC and intra-H.261 are not overly affected by data loss, because image errors do not propagate from frame to frame.

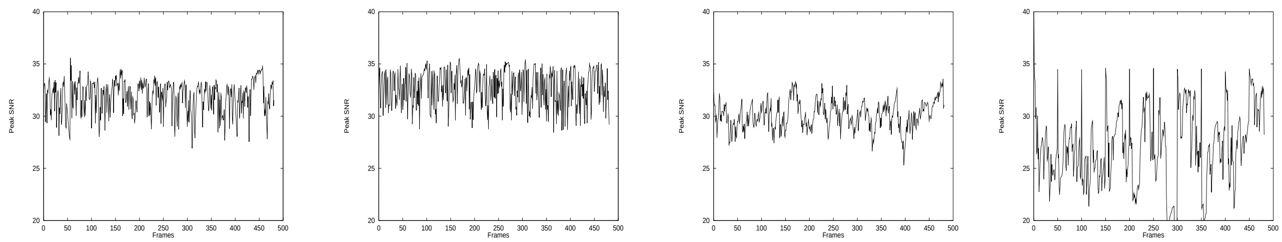
4 Conclusions

In this paper we presented RVC, a new hierarchical video coding scheme suitable for transmission over networks that provide various qualities of service, including guaranteed bandwidth and best-effort channels. Our objective is to minimize the amount of bandwidth reserved while utilizing the available best-effort bandwidth. However, the coding scheme has to compensate for the expected data loss from the best-effort channel to avoid image degradation. Some earlier schemes (H.261, MPEG) require high reserved bandwidth to maintain good image quality, because they do not compensate for lost data. Other schemes (Motion



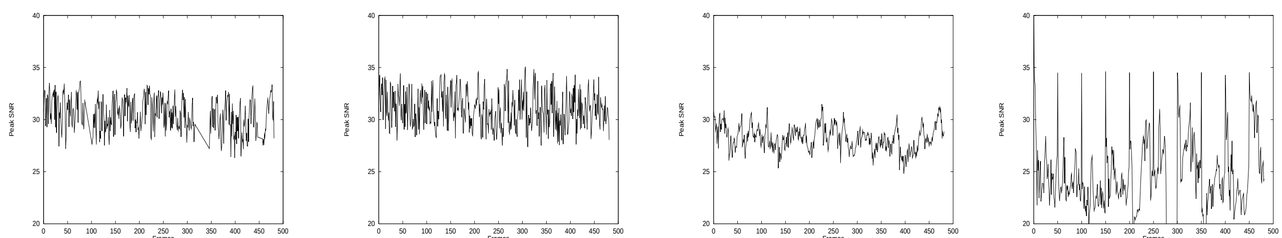
(a) RVC (avg = 33.47) (b) intra-H.261 (avg = 35.10) (c) H.261 (avg = 33.50) (d) TDC (avg = 33.41)

Peak signal-to noise ratio/frame under no data loss.



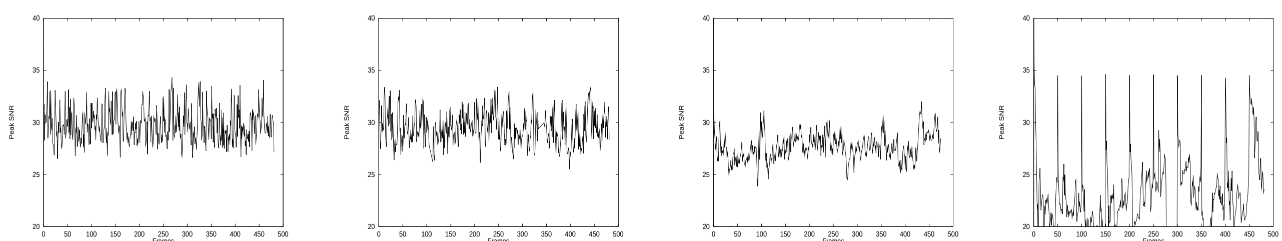
(e) RVC (avg = 31.79) (f) intra-H.261 (avg = 32.72) (g) H.261 (avg = 30.02) (h) TDC (avg = 26.70)

Peak signal-to noise ratio/frame under 10% data loss.



(i) RVC (avg = 30.53) (j) intra-H.261 (avg = 31.02) (k) H.261 (avg = 28.25) (l) TDC (avg = 24.38)

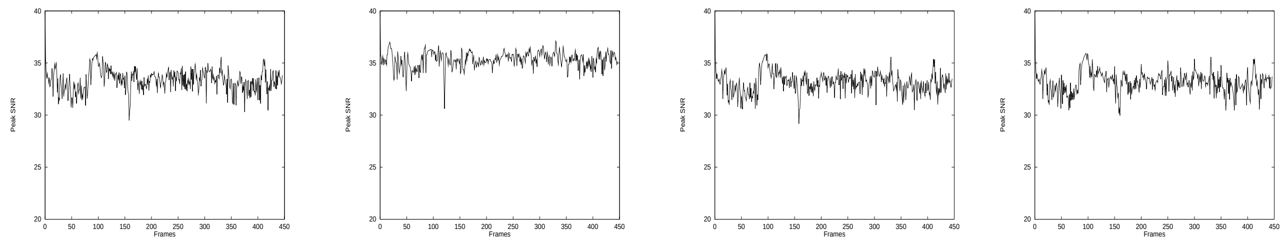
Peak signal-to noise ratio/frame under 20% data loss.



(m) RVC (avg = 29.44) (n) intra-H.261 (avg = 29.80) (o) H.261 (avg = 27.55) (p) TDC (avg = 22.56)

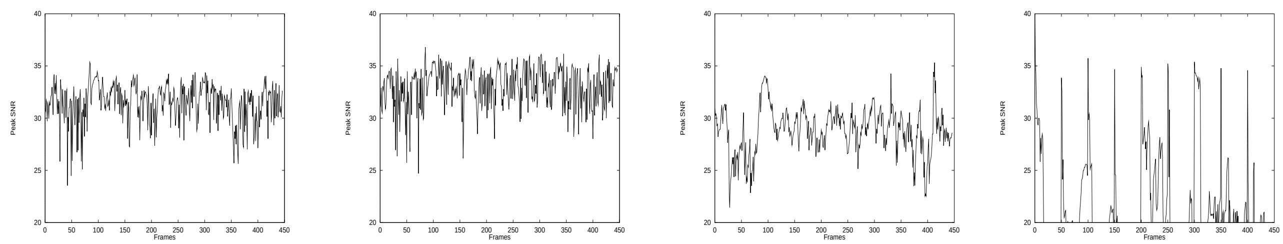
Peak signal-to noise ratio/frame under 30% data loss.

Figure 5: Peak signal-to-noise ratio results on low-motion sequence.



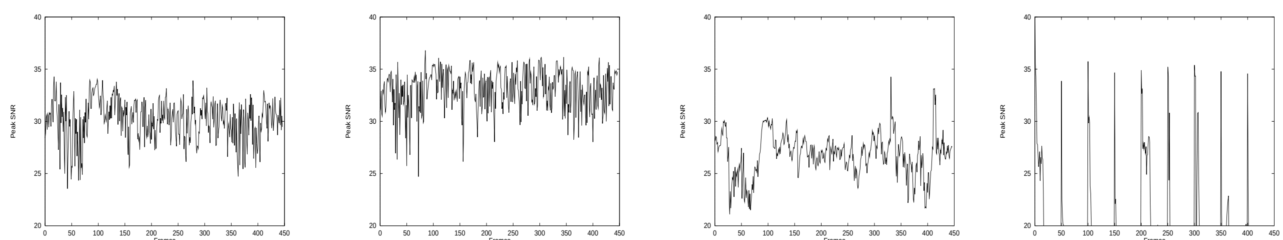
(a) RVC (avg = 33.36) (b) intra-H.261 (avg = 35.32) (c) H.261 (avg = 33.14) (d) TDC (avg = 33.18)

Peak signal-to noise ratio/frame under no data loss.



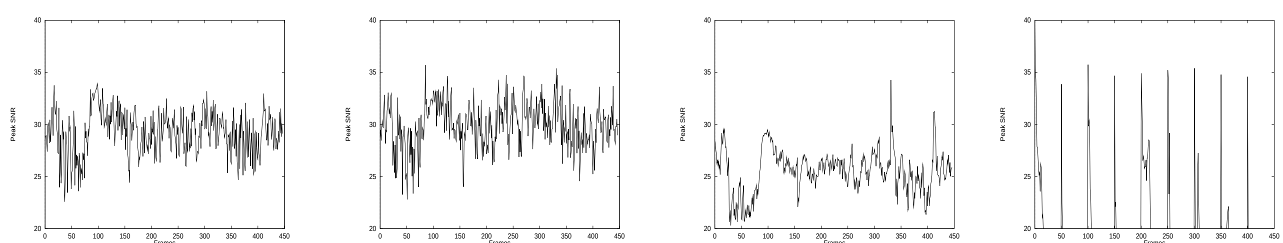
(e) RVC (avg = 31.44) (f) intra-H.261 (avg = 33.01) (g) H.261 (avg = 28.83) (h) TDC (avg = 21.15)

Peak signal-to noise ratio/frame under 10% data loss.



(i) RVC (avg = 30.14) (j) intra-H.261 (avg = 31.20) (k) H.261 (avg = 26.74) (l) TDC (avg = 18.09)

Peak signal-to noise ratio/frame under 20% data loss.



(m) RVC (avg = 29.17) (n) intra-H.261 (avg = 29.92) (o) H.261 (avg = 25.47) (p) TDC (avg = 16.37)

Peak signal-to noise ratio/frame under 30% data loss.

Figure 6: Peak signal-to-noise ratio results on high-motion sequence.

JPEG, intra-H.261) do compensate for data loss. However, they tend to increase the total data rate.

The RVC coding scheme shows a good tradeoff between data rate and tolerance of data loss. RVC's overall data rate is only slightly higher than H.261, and it exhibits excellent tolerance to data loss. Therefore, under the RVC scheme, the amount of reserved bandwidth can be minimized. Further, RVC's total data rate is up to 20% less than that of intra-H.261, while maintaining comparable image quality under data loss.

References

[1] E. Amir, S. McCanne, and M. Vetterli. A layered dct coder for internet video. *Proceedings of the IEEE International Conference on Image Processing*, September 1991.

[2] L. Delgrossi, C. Halstrick, D. Hehmann, R. G. Herrtwich, O. Krone, J. Sandvoss, and C. Vogt. Media scaling in a multimedia communication system. *Multimedia System*, 2:172–180, 1994.

[3] A. Eleftheriadis and D. Anastassiou. Meeting arbitrary QOS constraints using dynamic rate shaping of coded digital video. *Proceedings of the International Workshop on Network and Operating System Support for Digital Audio and Video*, pages 95–106, April 1995.

[4] CCITT Recommendation H.261. Video codec for audiovisual services at $p \times 64$ kbit/s, 1990.

[5] S. McCanne and V. Jacobson. vic: A flexible framework for packet video. *Multimedia '95*, pages 511–522, November 1995.

[6] I. Rhee, R. Rammohan, S.-Y. Cheung, and V. Sunderam. Achieving fairness and scalability in multicast video distribution using receiver-only rate control. Submitted to *ICDCS '97*, 1996.

[7] The ATM Forum. Traffic management specification, version 4.0, February 1996.

[8] T. Turletti and J.-C. Bolot. Issues with multicast video distribution in heterogeneous packet networks. *Proceedings of the Sixth International Workshop on Packet Video*, September 1994.

[9] L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala. RSVP: A new resource ReSerVation protocol. *IEEE Network*, September 1993.