Error Resilience Support in H.263+

Stephan Wenger
Gerd Knorr
Technische Universität Berlin, Germany
{stewe, kraxel}@cs.tu-berlin.de
Jörg Ott
Universität Bremen, Germany
jo@tzi.uni-bremen.de
Faouzi Kossentini
University of British Columbia, Canada
faouzi@ece.ubc.ca

Abstract: *The version 2 of the ITU recommendation H.263, better known as H.263+, includes a number of new mechanisms to improve coding efficiency and support various types of networks more efficiently. This paper provides an overview of the error resilience optional modes of H.263+ and describes the use of such modes in various network scenarios.*

1 Introduction

In the past, most video compression and coding standards were developed with a specific application and networking infrastructure in mind. For example, the ITU-T recommendation H.261 [1] was optimized for use with interactive audiovisual communication equipment, e. g. a videophone, and in conjunction with the H.320 series of recommendation as multiplex and control protocols on top of ISDN [2]. Consequently, the H.261 designers made various design choices that limit the applicability of H.261 to this particular environment.

The ITU-T Recommendation H.263 Version 2 — in this paper abbreviated as H.263+ and now ratified by the ITU-T [3] — is the very first international standard in the area of video coding which is specifically designed to support the full range of both circuit-switched and packet-switched networks. H.263+ contains functionalities that improve the quality of video transmission in error-prone environments and non-guaranteed quality of service (QoS) networks.

In this paper, we first present a short description of the coding structure specified by H.263+, and then, in some detail, H.263+'s optional error resilience oriented modes. We will next introduce both the networks and the transport hierarchies whose characteristics were used in our simulation work. Although there are many combinations of widely used networks and protocol hierarchies, we will focus on five scenarios based on wired networks[1] with different network characteristics. The selected scenarios cover a large part of today's infrastructure for multimedia communication. For each of these scenarios, we recommend and theoretically justify a set of combinations of the error resilience oriented H.263+ optional modes, that have proven to be

[1] Note that, although H.263+'s error resilience modes can be beneficial for wireless networks, they are mostly designed for wireline networks, the only widely deployed non-LAN networks that support interactive, low bit rate video communications at an acceptable quality. As wireless networks that support this type of video communications become available, ITU-T will likely provide optimized error resilience support for such networks.

effective in our research work. Corresponding simulation results are then presented. Conclusions are drawn in the last section.

2 Background

The following section contains background information on H.263+ and the relevant networks and transport protocol hierarchies

2.1 H.263+: An overview

In this section, we will provide a brief overview of H.263+. We will begin with a short outline of the standard's history and future. After we providing information on H.263+'s video coding structure we describe those optional modes, which are intended to improve video coding efficiency. The reader may refer to [4] for a more detailed description to such modes as well as a discussion of their functionalities and achievable coding efficiency levels.

2.1.1 ITU-T's family of Video Coding recommendations

The expert groups of the ITU-T have developed a series of video compression/coding recommendations, as have done other standardization bodies, namely ISO/IEC. The recommendation H.120, which was finalized in 1988, introduced the concepts of inter picture prediction and the Discrete Cosine Transform (DCT) in a standard document. Later, systems based on the recommendation H.261 had significant success in the marketplace. Video coding methods complying with this ITU-T recommendation are still the most often used coding methods for interactive applications like videophone and videoconferencing systems.

H.263 was developed mainly to improve the coding efficiency as compared to H.261, thus allowing lower bit rates than p x 64 kbit/s (as used in modem-based communication) to be used while still maintaining acceptable quality video. H.263 has had some success in the marketplace, in both PSTN videotelephone and ISDN-based high-end videoconferencing systems.

During the standardization process of H.263, the ITU-T decided to adopt the MPEG-2 video coding standard (ISO/IEC 13818-2) as the recommendation H.262, targeting the high bit rate, high quality market.

In 1996, ITU-T decided to continue the development of H.263 to achieve improved compression performance and to better support the upcoming multimedia communication systems based on non- guarantied QoS packet networks. The working name of the subject project is H.263+, and this name for the now ratified standard, which is officially known as H.263 (1998), is well accepted in both academia and industry.

2.1.2 H.263+ Baseline Operation

H.263+, not using any of its optional modes and mechanisms is called the H.263+ baseline. The H.263+ baseline is identical to the H.263 Version 1 baseline. Its operation is base on motion compensation based inter picture prediction to reduce temporal redundancies and DCT-based transform coding to reduce spatial redundancies in the prediction difference pictures. DCT-based transform coding consists of forward DCT, quantization of the resulting DCT coefficients, and Huffman-based variable length coding.

A H.263+ baseline encoder expects YUV 4:1:1-coded, non-interlaced video data as input. Five input resolutions are defined: SQCIF (128 x 96), QCIF (176 x 144), CIF (352 x 288), 4CIF (704 x 576) and 16CIF (1408 x 1152). Only SQCIF and QCIF are mandatory for every decoder.

This input picture is divided into macroblocks consisting of four luminance blocks and two chrominance blocks of 8x8 pixel each. For picture formats up to CIF, one line of macroblocks is called a Group of Blocks (GOB). For the 4CIF format, a GOB consists of 2 lines of macroblocks, and for the 16CIF format, a GOB consists of 4 lines of macroblocks. GOB headers are used to allow re-synchronization in case of decoding errors and can therefore be assumed as only syntactical means of dividing the entropy coded bit stream.

Most H.263+ baseline encoders perform motion estimation for each macroblock, although this is not mandated by the standard. For each macroblock, the encoder decides on using inter coding (motion compensated prediction and DCT coding of prediction difference blocks) or intra coding (DCT coding of the original blocks). The decision process is also not defined in the recommendation, although it is usually based on a heuristic use of the results of the motion estimation algorithm. In both the inter coding and intra coding cases, the DCT is applied to each of the six 8x8 constituent blocks of the difference macroblock (inter) or the original macroblock (intra). For each 8x8 block, the resulting DC coefficient and 63 AC coefficients are quantized and, along with applicable motion vectors, are Huffman-based variable length coded. Finally, a rate control method is used to decide on the quantization step size and the type of prediction information. Although several rate control methods are suggested in test model documents, none is specified in H.263+.

The aforementioned H.263+ baseline operation is applied to each macroblock of a picture. After a complete picture is coded, the same output data is used to reconstruct a reference picture, which is supposed to be identical to the picture at the decoder, assuming an error-free environment. This reference picture is used for motion compensated prediction.

2.1.3 H.263 Version 1 Optional Modes

To provide different tradeoffs between coding efficiency and complexity, H.263 version 1 includes four optional coding modes. The use of these modes is signaled in the header of a H.263 coded picture. Using optional modes may change significantly the syntax and semantics of a H.263 baseline bit stream and some of the modes may add additional codepoints to the bit stream. We next provide a summary of the four optional modes.

2.1.3.1 Annex D: Unrestricted Motion Vector mode

The unrestricted Motion vector mode extends the range of each of the two motion vector components from +/-16 to +/-32. This mode is especially useful in the case of camera pan and for video scenes with high motion variations.

2.1.3.2 Annex E: Arithmetic Coding mode

This mode replaces the Huffman-based variable length coding (VLC) of DCT coefficients and motion vectors with arithmetic coding. As the VLC and the arithmetic coding are both loss-less, using arithmetic coding will usually lower the bit rate (although typically by less than 5%) without impacting video reproduction quality. Equivalently, if the bits saved by using arithmetic coding are used to further reduce distortion, a small increase in quality then results. The cost, however, is a significant increase in computational complexity. This, coupled with IPR-related problems associated with arithmetic coding, continues to prevent widespread use of this mode.

2.1.3.3 Annex F: Advanced Prediction mode

The Advanced Prediction mode allows the coding of four motion vectors per macroblock, instead of only one motion vector per macroblock in case of baseline. These four motion vectors correspond to the four constituent luminance 8x8 blocks. Using the Advanced Prediction mode improves the picture quality significantly with only a typically small increase in bit rate. However, this mode also requires additional complexity, especially at the encoder.

2.1.3.4 Annex G: PB-frame mode

The PB-frame mode increases significantly the coding efficiency by doubling the video frame rate with only a moderate increase in bit rate. This is achieved by coding two source pictures as one PB-frame. The second source picture is coded as in the case of B-pictures used in MPEG-1/2, but without backward prediction using the subsequent P-picture.

2.1.4 H.263+ Optional modes and enhancements

With the exception of the Unrestricted Motion Vector mode, H.263's optional modes are essentially supported, both in syntax and in semantics, by H.263+. The latter also supports twelve new optional modes as well as an enhanced Unrestricted Motion Vector mode. Furthermore, the enhanced picture header can carry additional information beyond the signaling of the new optional modes. Virtually all mode permutations are allowed, and a few exceptions that clearly marked in the H.263+ standard. For example, it is forbidden to use the scan-order sub-mode of the Slice Structured mode in conjunction with the Independent Segment decoding mode. To ease the implementation of H.263+ decoders and to speed up the capability exchange process between H.263+ capable terminals, the ITU-T recommends a set of "preferred mode combinations" in a non-normative Appendix of this standard. Next, we will describe briefly the H.263+ enhancements and non-error resilience oriented modes.

2.1.4.1 The H.263+ Enhanced Picture Header

In H.263+, the picture header allows the signaling of 1) the new optional modes, 2) new picture types for the Temporal, Spatial, and SNR Scalability mode, 3) custom picture sizes, which can be as large as 2048 x 1152 pixels and can vary by as few as 4 x 4 pixels, 4) a custom picture clock and 5) a custom pixel aspect ratio. The picture header can be either conveyed completely, or in one of several abbreviated forms. This feature is necessary due to the possibility of requiring large header sizes, which can reach several 100 bits. Abbreviated picture headers improve the coding efficiency but they sometimes have a negative impact on error resilience. If the H.263+ enhanced picture header is present, some mechanisms of the standard work slightly different from those of H.263 version 1. An example is the rounding of coefficient data during quantization.

2.1.4.2 Annex D: Unrestricted Motion Vector mode

If the H.263+ enhanced picture header signals the use of Annex D, the motion vector components may have any size, up to the size of the picture. A new syntax is introduced to allow for the coding of large-component motion vectors. Although it does not restrict the motion vector range, H.263+ suggests that the maximum range be negotiated by external means to facilitate decoder implementation and resource allocation.

2.1.4.3 Annex I: Advanced Intra Coding mode

The Advanced Intra Coding mode improves the compression efficiency of I-pictures or P-pictures with a large percentage of intra coded blocks. The performance gain is achieved by spatial prediction of DCT coefficient values of intra coded blocks.

2.1.4.4 Annex J: Deblocking Filter mode

An adaptive filter inside the coding loop is here used to reduce the amount of blocking artifacts in the reproduced video pictures by filtering across block boundaries. This often improves the video reproduction quality.

2.1.4.5 Annex L: Supplementary Enhancement Information Specification

Annex L defines various codepoints for supplementary information. Decoders do not have to be aware of this type of information. Thus, such information does not change the semantics, or prevent the correct decoding, of the bit stream for decoders not implementing this mode[2]. Corresponding decoders, however, can benefit from supplementary information as they can therefore include features s8uch as various picture freeze and release commands, video chroma keying, and tagging information for external picture-exact synchronization.

2.1.4.6 Annex M: Improved PB-Frame mode

The PB-frame mode defined in Annex G does not allow truly bi-directional prediction, which is now provided by Annex M. This mode offers a significant advantage over Annex G's PB-frame mode in terms of coding efficiency and robustness, while requiring little additional computational complexity. The annex can be viewed as a substitution of Annex G, but was added as an additional feature to allow downward compatibility.

2.1.4.7 Annex O: Temporal, Spatial, and SNR Scalability mode

The use of the Temporal, Spatial, and SNR Scalability mode yields a H.263+ layered bit stream. A base layer can be augmented by several enhancement layers, which improve the temporal and/or spatial resolution, and/or video reproduction quality. Enhancement layers can use hierarchically lower enhancement layers or the base layer as their anchors (i. e. for reference). However, temporal enhancement layers are not allowed as be used references for any form of prediction. In real-world systems, the base and enhancement layers can be either multiplexed into a single bit stream (for which H.263+ provides the codepoints in the enhanced picture header), or conveyed over individual transport channels.

Temporal scalability is achieved by means of B-pictures, as in MPEG 1/2. Using B-pictures will often increase the coding efficiency, but normally has a negative effect on latency. If spatial scalability is used, the lower layer of a picture is up-sampled by a factor of two in each dimension before it is used as the reference. This allows the very efficient predictive coding of larger pictures, if a small base picture is already available at the decoder. SNR scalability is similar to spatial scalability, with the exception that the reference picture is not up sampled.

[2] Supplementary Enhancement Information is carried by the PEI/PSUPP mechanism, which was reserved for future use in H.263 version 1, but nevertheless already present in the syntax.

2.1.4.8 Annex P: Reference Picture Resampling mode

This mode allows the resampling of the previous reference picture prior to its use as a reference. The mode enables effective global motion compensation (useful in cases of camera pan) and predictive dynamic resolution conversion. Apart from this, special effects such as warping of images are also possible.

2.1.4.9 Annex Q: Reduced Resolution Update mode

The Reduced Resolution Update mode allows the temporally use of lower resolutions on well-defined spatial areas of the picture, while maintaining the higher resolution on the rest of the picture. The mode is helpful for the efficient coding of local areas of heavy motion while still achieving a good picture quality and high resolution for the rest of the picture.

2.1.4.10 Annex S: Alternative INTER VLC mode

This mode improves coding efficiency for some sequences by using a different VLC table for inter coded blocks. It has been proved useful especially for high quality / high bit rate coding.

2.1.4.11 Annex T: Modified Quantization mode

The Modified Quantization mode enables the encoder to choose one of several quantizer tables on a per macroblock basis. This mode can improve coding efficiency by reducing chroma artifacts through reducing the step size during quantization of the chrominance component of the video signal. This mode also increases the range of representable coefficient values in case of numerically small quantizer values.

2.2 Error-Resilience optional modes of H.263+

Among the twelve new optional modes, three modes are designed for the efficient support of the various network types with their different congestion characteristics. Two of the three modes are quite complex and contain various sub-modes. The third mode is simple in both concept and implementation. One of the optional modes of H.263 version.1 is also covered here, because it handles error resilience related issues. All four error resilience oriented modes of H.263+ are described, in detail, in the following sections.

2.2.1 Annex H: Forward Error Correction mode

The Forward Error Correction (FEC) mode is the oldest of the error resilience oriented optional modes. If Annex H is used, the H.263 bit stream is divided into FEC-frames of 492 bits each. A 19 bit BCH forward error correction checksum is calculated for all the bits of such a FEC-frame, along with one bit that is necessary for the synchronization to the frame structure. This FEC coding allows the correction of single bit errors in each FEC-frame and the detection of two bit errors for an approximately 4% increase in bit rate. The use of BCH-FEC requires, however, that all bits be transmitted because even one missing bit would require a complete re-synchronization to the FEC-frame structure. Such a re-synchronization may take about 30,000 bits of data (roughly a quarter of a second at a speed of 128 kbit/s) and usually impairs predictive coding significantly because several pictures might get lost during the re-synchronization process. The FEC mechanism of Annex H is designed for ISDN, which is an isochronous, very low error rate network.

2.2.2 Annex K: Slice Structured mode

The Slice Structured Mode replaces the original Group of Block (GOB) structuring of the macroblocks of a picture by a slice structure. Slices consist of a number of macroblocks belonging to the same picture. These macroblocks might be arranged either in scanning order as in MPEG 1's slices [5], or in a rectangular shape. In both cases, any macroblock of a picture belongs to exactly one slice. All macroblocks of one slice can be decoded independently from the content of other slices, because no dependencies such as prediction of motion vectors are allowed across slice boundaries. There is, however, the need to have the information of the picture header available to decode a slice, because the information conveyed in the picture header is not copied to the slice headers. The scan order slices sub-mode is often useful if small packet sizes are used, whereas the rectangular slices sub-mode is helpful in achieving packet loss resilience and low codec delay at higher bit rates. Each of the two sub-modes can be used either with a fixed scan-order transmission of the slices or following an arbitrary order. The latter one makes decoder implementation more difficult but minimizes latency in lossy environments. The former one is more appropriate for heavily pipelined hardware architectures, which might not allow random decoding of data.

2.2.3 Annex R: Independent Segment Decoding mode

The Independent Segment Decoding mode enforces the treatment of segment boundaries as if they were picture boundaries. A segment is defined as a slice, a GOB, or a number of consecutive GOBs with empty GOB headers. This mode allows the independent decoding of picture parts, if and only if, the shape of the independently decodable segments remains identical between two I-pictures. In such a case, the motion-compensation related import of previously corrupted picture data outside the segment boundaries during the reconstruction process could be avoided. The Independent Segment Decoding mode can be used for special effects like spatial video mixing, but it can also achieve error resilience by eliminating error propagation between well-defined spatial parts of a picture.

2.2.4 Annex N: Reference Picture Selection mode

The Reference Picture Selection mode allows the use of an earlier than the last transmitted picture to serve as the reference picture for inter picture prediction. It is also possible to apply the Reference Picture Selection mode to individual segments rather than full pictures. The temporal reference (TR) of the to be used reference picture is conveyed in the picture/segment header to inform the decoder which of its several reference pictures should be used.

The Reference Picture Selection mode may be used with or without a back channel. If a back channel is used, it can be either multiplexed onto the H.263+ data stream of the opposite direction (VideoMux back channel sub-mode), or conveyed out of band (separate logical channel sub-mode). The VideoMux back channel sub-mode is only applicable for bi-directional video communication, because the back channel messages are conveyed within the video data of the opposite direction. The separate logical channel sub-mode is, from a software engineering point of view, often seen as a more friendly way of conveying back channel data, but makes a separate data channel necessary, which incurs significant additional overhead in some environments.

Back channel messages contain positive or negative (or both positive and negative) acknowledgments of a decoded picture along with the TR of the picture. By using this information, the encoder can keep track of the last correctly decoded picture at the decoder. Once the encoder learns about a non-correctly decoded picture at the decoder through a back channel message, it can react accordingly by using a known-as-correct reference picture (often

the most recent one) for further prediction. Since the number of stored reference pictures is limited and those reference pictures are usually kept in a sequential order, it might even be necessary for the encoder to react by sending a full I-picture, if it observes that no reference picture is present at the decoder. In the case of point-to-point connections with low transmission delay characteristics, back channel mechanisms are a valuable addition to achieve temporal error resilience, especially if augmented by error concealment techniques [6, 7].

In application scenarios that involve some ten, hundred, or even more endpoints and that are based upon multicast or broadcast communication mechanisms, as available in LANs, Intranets, and the Mbone of the Internet, back channels are obviously not applicable. For those scenarios, Annex N's sub-mode without a back channel can be used instead. One possible method of using the Reference Picture Selection mode without back channel is known as Video Redundancy Coding (VRC) [8]. VRC can be used in conjunction with the spatial error resilience mechanisms of Annex R and Annex K to achieve spatio-temporal error resilience.

The principle of the VRC method is to divide the sequence of pictures into two or more *threads* in such a way that all camera pictures are assigned to one of the threads in a round-robin fashion. Each thread is coded independently. Obviously, the frame rate within one thread is much lower than the overall frame rate: half in case of two threads, a third in case of three threads and so on. This leads to a substantial coding penalty because of the generally larger changes and the longer motion vectors typically required to represent accurately the motion related changes, between two P-pictures within a thread. In regular intervals, all threads converge into a so-called *Sync frame*. From this Sync frame, a new thread series is started. Figure 1 illustrates VRC with two threads and 3 frames per thread.

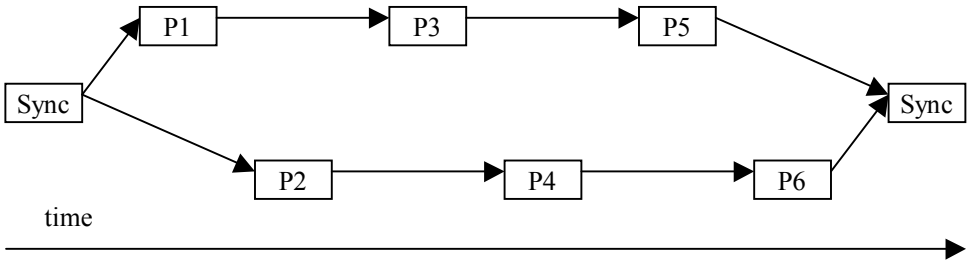

Figure 1: VRC with 2 threads and 3 frames per thread

If one of these threads is damaged because of a packet loss, the remaining threads stay intact and can be used to predict the next Sync frame. It is possible to continue the decoding of the damaged thread, which leads to slight picture degradation, or to stop its decoding which leads to a drop of the frame rate. If the length of the threads is kept reasonably small, however, both degradation forms will persist only for a very short time, until the next Sync frame is reached. Figure 2 illustrates the workings of VRC when one of the two threads is damaged.

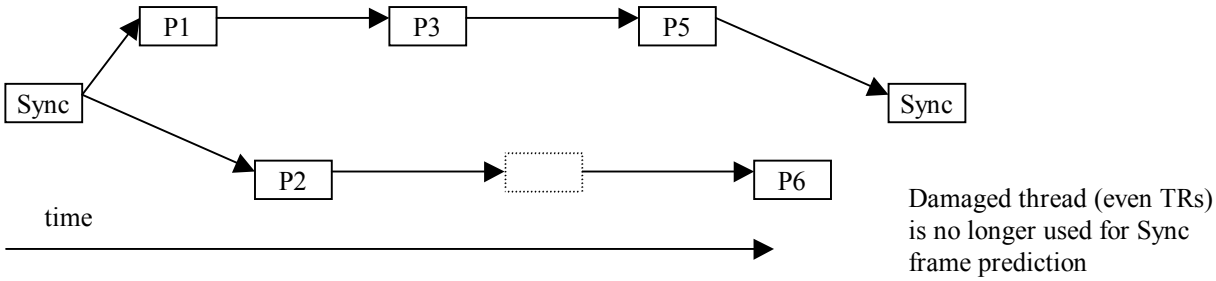

Figure 2: Frame loss with VRC

Sync frames are always predicted out of one of the undamaged threads. This means that the number of transmitted I-pictures can be kept small, because there is no need for complete re-synchronization. Only if all threads are damaged between two Sync frames, a correct Sync frame prediction is no longer possible. In this situation, annoying artifacts will be present until the next I-picture is decoded correctly, as it would have been the case without employing VRC.

2.3 Networks for Multimedia Communication

In this paper, we set our focus on interactive communication over well-established network structures. In the rest of the paper, we cover the following networks:

- PSTN: the Public Switched Telephony Network is available virtually everywhere in the world. Today's modem technology offers net bit rates of about 28 kbit/s, which allows a low-quality video transmission, amended with telephony quality audio. The whole standardization process of H.263 and H.324 was triggered by the appearance of non standardized PSTN video telephones in 1994.
- ISDN: although not available everywhere, the Integrated Services Digital Network is today's most widely used circuit switched network for Interactive multimedia Communication. Videotelephone and Videoconferencing systems based on the H.320 family of ITU-T recommendations are still the only affordable, medium to high quality, videoconferencing solutions for most business users
- The Internet: today's largest and most popular data network can also be used for multimedia communication. The Internet cannot easily be compared with PSTN and ISDN, because it is not a circuit switched, but a packet switched network. Generally, the Internet offers variable bandwidth, variable transmission delay and variable error characteristics depending on the general conditions of the Internet backbone and the dialup connections to that backbone (if used). Its characteristics are determined by both the involved dialup-links and the backbone. An overview of the multimedia communication protocols used on the Internet, as well as a detailed discussion of their characteristics, can be found in [9].

A more comprehensive overview of networks for multimedia communications, which also covers all the protocols hierarchies discussed below in some detail, can be found in [10].

2.4 Protocol Hierarchies

For all three aforementioned networks, the relevant standardization bodies defined at least one protocol hierarchy for multimedia communication. In case of PSTN and ISDN, the ITU-T is the relevant standardization body, whereas in case of the Internet, the Internet Engineering Task

Force (IETF) is the accepted authority [11, 12]. However, the ITU-T has also developed a protocol family, originally intended only for Local Area Networks (LANs), but is today also used in the context of the Internet for both Internet telephony and Internet multimedia conferencing.

2.4.1 H.324 for PSTN

If the telephony network is to be used for multimedia conferencing, most commercial products rely on the H.324 protocol hierarchy to insure interoperability. The ITU-T recommendation H.324 entitled "Terminal for low bit rate Multimedia Communication" [13], provides an overview of PSTN multimedia terminals and references all other ITU-T recommendations which are necessary to build such a terminal in a standard conformant way. For this reason, the standardization community refers to a H.324 family of recommendations.

Since PSTN is a circuit-switched, point-to-point network, only point-to-point communication relationships are possible without the means of special networking equipment, which, as of today, is neither standardized nor available. This is of critical importance when discussing error resilience, because some of the error resilience mechanisms are only applicable for the point-to-point case.

Out of the various ITU-recommendations of the H.324 family, only the definition of the low-level transport is relevant to this paper. This definition can be found in the ITU-T recommendation H.223 entitled "Multiplexing Protocol for Low Bit rate Multimedia Communication" [14]. Three different types of Adaptation Layers (ALx) are available which have different characteristics in terms of error probability and delay (since low delay channels do allow higher error rates, whereas reliable channels might have indefinitely long delays). While AL1 and AL2 serve different duties, AL3 is designed for the use with coded video. Video data is encapsulated in small, variable length packets (typically some 100 bytes, although larger packet sizes can be negotiated). A 16 bit CRC for each packet allows error detection. The packetization overhead for each packet is 1 to 3 bytes, plus the error control information of AL3. AL3 includes an optional retransmission protocol, which sometimes allows the retransmission of a lost or corrupted packet.

The retransmission of AL3 relies on the fast arrival of the confirmation messages, which give indications about correctly transmitted packets. Those confirmation messages arrive at the sender of the original message with twice of the one-way delay, since a complete round-trip of data and confirmation is necessary. The retransmission of the damaged packet after notification will incur a third one-way delay, resulting in three times a one way delay. If such a one-way delay is already high – which can be due to intercontinental satellite connections or to the system design – then the AL3 retransmission will add such a substantial delay that it's use is no more acceptable in interactive applications. For such cases, which are possible although not likely as shown during Interoperability-events of the IMTC [15], additional H.263+ error resilience support would be helpful.

2.4.2 H.320 for ISDN

Videotelephone and Videoconferencing systems using ISDN are the most widely used multimedia communication systems in the market today. Their characteristics are defined in the ITU-T recommendation H.320 entitled "Narrow-band visual telephone systems and terminal equipment" and in various other recommendations referenced in H.320 [2]. In this paper, we will only provide an overview of the low-level transport protocol, which is defined in the ITU-T recommendation H.221 entitled "Frame structure for a 64 to 1920 kbit/s channel in audiovisual teleservices" [16].

H.221 is the multiplex and bonding protocol for H.320-terminals. Up to 30 ISDN-B-channels can be bundled together to form a "super channel" with a bit rate of n x 64kbit/s. The media channels for audio and video, and the data information along with a "service channel" for control information are multiplexed onto the "super channel". For audio and video information, H.221 does not perform any error control, but relies completely on the error resilience of the media coding – which because of ISDN's isochronous nature and its low error rates, is possible. The protocol offers only a bit-oriented, unprotected, point-to-point transport service.

In contrast to PSTN systems, multipoint capable ISDN systems are both standardized, and commercially available in form of Multipoint-Control-Units. These units may or may not contain transcoding devices for spatially mixing video data. If a MCU with transcoding (sometimes called video-mixing MCU) is used in a multipoint scenario, none of the point-to-point only error resilience mechanisms is applicable.

2.4.3 H.323 for LANs and the Internet

Multimedia communication over packet oriented networks, such as Local Area Networks (LANs) and the Internet, are defined in both Internet-RFCs and in the H.323-series of ITU-T recommendations. The recommendation H.323 entitled "Visual telephone systems and equipment for local area networks which provide a non guaranteed quality of service" and its associated protocol support, focus on small, closely coupled and highly interactive conference scenarios, and leave the large, loosely coupled conferences to the "native" Internet protocols described in the next section. Again, among the numerous recommendations of the H.323 family, only the low-level transport recommendation H.225 entitled "Media stream packetization and synchronization on non-guaranteed quality of service LANs" will be considered in this paper [17].

H.225 defines the transport/multiplex layer. Each data or media stream is transported in its own virtual transport stream. For non-real-time data streams, an HDLC-type protocol is used to ensure reliable transport. Real-time media data, however, is transported by means of the Real-time Transport Protocol RTP [18]. Each RTP packet contains a minimum of 40 bytes header information[3]. In order not to waste bandwidth, the relationship between payload data and packetization overhead should be optimized, leading to large payload sizes and consequently to large packets. The effective maximum size (often abbreviated as MTU, maximum transfer unit) of a packet varies from network to network and is sometimes smaller than the maximum packet size allowed by the underlying protocols. In the Internet, as well as in most IP-based LANs, packet sizes of less than 1500 bytes have been proven to yield good performance [19, 20, 21, 22], although sizes up to 64 KB are allowed by the protocols.

On Internet connections using dialup links (modem or ISDN), the bit rate for video is strictly limited. If, however direct connections to the Internet backbone are used, the link is usually shared between many users. In such a case, it is often possible to overdraft the bit rate budget allocated for video for a short period of time (e. g. to send an I-picture). As we will later see, this is important for error resilience, because some error resilience mechanisms require unusual large pictures, whereas others do not.

[3] RTP uses the Unreliable Datagram Protocol UDP to transport its Protocol Data Units (PDUs). UDP relies on IP. The header of an RTP packet consists of 16 byte RTP header, 8 byte UDP and 20 byte IP header. RTP payload specific headers might add additional overhead.

2.4.4 Native Internet Multimedia Conferencing Protocols

H.323, although multipoint-capable, sets its focus on highly interactive, but small conferences. The IETF multimedia conferencing approach, in contrast to the H.323 one, is designed primarily for medium to large groups of hundreds or even thousands of participants, which are much less interactive (thus allowing higher latency times). In such an approach, two-way communication mechanisms such as retransmissions or acknowledgments of correctly transmitted video packets via back channel messages should not be used. The protocols used in a native Internet environment are very similar to those of H.323. For the transport of media data, RTP is used on top of UDP/IP. Any higher functionality is handled by protocols like the Session Announcement Protocol SAP [23] and the Session Initiation Protocol (SIP) [24]. Both offer a very limited amount of capability control, which might be enhanced by the Simple Conference Control Protocol SCCP [25]. Finally, note that the multicast communication of native Internet conferences on the Mbone [26] offers a conducive environment for the development and use of layered codecs.

2.5 Characteristics of the network/protocol combinations

Having identified three networks and four corresponding protocol environments, we provide, in **Error! Reference source not found.**, some important characteristics of the network/protocol combinations, expressed in terms of bandwidth, bit error probability, typical packet size and packet loss probability.

Network/ Protocol	**Bandwidth kbit/s**	**Bit-Error Probability**	**Typical Packet-Size (bytes)**	**Packet loss probability**
PSTN, H.324	<=20, variable	low (bit errors can be identified AL3) see details in Section 3.1	Small (e. g. 100 bytes)	0% to 5%, depending on line and QoS and allowed delay (retransmission mechanism of AL3) details in Section 3.1
ISDN, H.320	n x 64, fixed	$< 10^{-8}$	N/A	N/A
Internet, H.323, Modem-Dialup	<=20, variable	0	Up to 1500, smaller, if coded I-pictures are smaller	0% to 100% (line-quality and backbone-conditions)
Internet, H.323, ISDN dialup or leased line	>= 64, variable	0	1500 bytes	0% to 100% backbone-conditions

Table 1: Network/Protocol combination characteristics

Based on the above table, we next develop the following five scenarios. Such scenarios are intended to represent most of today's multimedia conferencing environments.

3 Five different Networking scenarios

Having the various networks and their corresponding protocols in mind, we propose a framework which includes the following five scenarios:

3.1 Scenario 1: PSTN, H.324

H.324-based systems are free to negotiate protocol parameter values such as packet size and allowed round trip delay. Such values, and especially the observed round trip delay, impact substantially the performance of the error resilience modes. To minimize packetization delay, packets with small sizes are usually negotiated. A typical payload size for AL3 video packets in H.324-systems is 120 bytes, a number often used in various interoperability tests (performed in the context of the IMTC) of commercial H.324 systems.

Moreover, the maximum delay of a H.223 AL3 channel has a significant impact on the effectiveness on AL3's retransmission algorithm. A low-delay system can use AL3's retransmission algorithm in a very effective way, thus reducing to nearly zero bit or packet loss errors. In such a case, AL3 will provide a practically error-free environment, making additional error resilience mechanisms unnecessary.

Many of today's H.324 systems, unfortunately, have a significant end-to-end delay, due mainly to the integration of H.324 protocol mechanisms within a PC operating system environment, which is usually not optimized for real-time applications. The typical end-to-end delay for video in such a system can be well above 0.5 seconds. Therefore, AL3 retransmission should be avoided if an interactive use and thus a reasonable delay is to be achieved. Since the special modem protocols used for H.324 do not perform any own error control or correction, significantly high bit error rates can occur. AL3's CRC checksum allows the easy detection of those bit-errors, even if the retransmission mechanism is disabled.

3.2 Scenario 2: ISDN, H.320

Due to the low error rates of ISDN (in the range of 10^{-8}) and its isochronous nature, error resilience mechanisms are almost unnecessary. As confirmed later by our simulation results ISDN bit error rates are too small to impact video reproduction quality.

3.3 Scenario 3: Internet, H.323, Modem-Dialup

PPP (for the dialup-part of the connection) and IP/UDP (for the backbone transmission) already provide a bit-error-free environment. Packet losses, both on the backbone and on dialup connections, are however still possible. The packet loss rate on the backbone is often small, because of the low amount of data traffic. On the other hand, the packet loss rate on the dialup part of the connection may be substantial. In this scenario and the subsequent ones, we will assume a maximum packet loss rate of 20%, because our experimental results have shown that higher packet loss rates render the reproduced video sequences simply useless. Additionally, although the bandwidth in the current scenario is very small, it should be noted that the Internet does provide Multicast transport mechanisms, which allow the same data to be transported to several users without sending it more than once. If this is the case, bi-directional error resilience mechanisms are inappropriate.

3.4 Scenario 4: LAN or Internet, H.323, Point-to-Point case

The Internet-protocols (IP/UDP/RTP) used for this type of connection provides a bit-error-free environment. However, packet loss rates can become substantial. As stated above, we will restrict packet loss rates to a maximum of 20%. To avoid the substantial bit rate increase resulting from the use of a small payload in conjunction with a header of a minimum of 40 bytes per RTP packet, the selected packet size should be as large as possible. In this scenario, a reasonable maximum packet size for the Internet is 1500 bytes. Due to a direct connection to the

Internet backbone, it is possible to overdraft the bit rate budget associated with the particular virtual connection from time to time (e. g. for an I-picture update). Finally, since the scope of this scenario is limited to the point-to-point case, any bi-directional mechanisms are acceptable.

3.5 Scenario 5: LAN or Internet, H.323 or native Internet protocols, multipoint-case

This scenario is similar to Scenario 4. Since multicast or broadcast communication might be used to reduce the network load, it is impossible to employ bi-directional error resilience mechanisms. All other remarks of Section 3.4 remain true.

4 Which modes for each scenario?

We now provide suggestions on how to apply H.263+'s error resilience modes to each of the above scenarios. The following recommendations are the result of a large number of simulation experiments, some of which are discussed in Section 5.

4.1 Mode combinations for scenario 1

A system environment with very low delay characteristics will allow the effective use of H.223 AL3's retransmission algorithm thus providing a nearly error-free environment. Other, especially PC-based systems, do not use AL3's retransmission today to keep the latency at an acceptable value. In both cases, systems will use small packets with sizes of around 120 bytes payload. For small packets, the scan-order slices sub-mode of the Slice Structured mode allows a very flexible fragmentation of a coded picture at the macroblock boundaries. Slices are independently decodable if the information of the picture header is available. Therefore, bit errors within a slice permit the decoding of the other slices of a picture, although it is still possible that artifacts present in the damaged picture parts will inevitably impact temporal prediction accuracy. Moreover, the overhead of using Slice Structured mode is typically well below 5%. Thus, we suggest that the Slice Structured mode be used for Scenario 1.

If the error probabilities values become significant, it is also helpful to use the Reference Picture Selection mode with a back channel (if not in a multicast situation). This mode introduces a very low additional bit rate overhead, especially if errors do not occur.

4.2 Mode combinations for Scenario 2

For Scenario 2, none of the optional modes, except the BCH Forward Error Correction mode (BCH-FEC) is necessary or helpful. This is due mainly to the high transmission quality of the ISDN network but also to the usually high bit rate overhead introduced by the other error resilience modes. BCH-FEC introduces a relatively low bit rate overhead of approximately 4%, making it suitable for the subject scenario. In the improbable case of loss of synchronization due to uncorrectable bit errors, the Full-Intra-Request mechanism of H.320's control protocol H.242 [27] can be used as a backup solution, allowing complete re-synchronization by requesting a new I- picture from the encoder.

4.3 Mode combinations for Scenario 3

Given the very low bit rate of approximately 20 kilobits/sec available for video, all parameters need to be well optimized. The packetization overhead of at least 40 bytes per packet makes it inevitable that exactly one coded picture be mapped into each packet, with the possible exception

of the usually much larger I-pictures. Consequently, the modes which allow more flexible splitting (Slice Structured mode) or independently decoding of parts of the picture (Independent Segment Decoding, mode) should not be used. However, the Reference Picture Selection mode, sub-mode with back channel, may be beneficial. This mode shows positive results in cases where the transmission latency of the transport between the two terminals is low (e. g, if both users have a dial-up connection to the same ISP). In such situations, a decoder can promptly update the encoder with the current packet loss condition and confer that an older reference picture be used for prediction. This is, however, only feasible if point-to-point communication is used.

4.4 Mode combinations for Scenario 4

Our research results have shown that using the Reference Picture Selection mode with a back channel outperforms all other error resilience mechanisms, provided that back channel messages can be conveyed within reasonable round-trip periods of time. Thus, if a back channel is available, such a mode should be here used. In addition, it may be helpful (especially for large picture sizes and high bit rates) to divide the picture into several independently decodable picture segments by applying the Slice Structure mode and the Independent Segment Decoding mode, and apply the Reference Picture Selection mode to such segments. Doing so will eliminate the propagation of coding artifacts from the subject slice to those that are outside its boundaries. The number and size of the rectangular slices should be chosen such that a very large percentage of all inter coded rectangular slices fit into one single packet. This will limit the typical coded slice size to about 1400 bytes, which is below the maximum packet size of 1500 byte (the latter number includes all header information, thus also the variable length H.263+ RTP payload header). The packetization method developed for H.263+, currently in the state of an Internet draft [28], enables the above type of packetization and recommends that the complete H.263+ picture header be added to the RTP payload header. This allows the decoding of the content of a packet even if a previous packet containing the picture header of the subject picture is lost.

4.5 Mode combinations for Scenario 5

In Internet/RTP environments, the actual packet loss rate is reported to the sender in the receiver reports of the Real-Time Control Protocol RTCP, which is defined within the RTP RFC [18]. This mechanism is always available, because RTCP is mandatory if RTP is used. This ensures that the necessary information about packet loss rates is available to the sender. Based on our experimental results, we here suggest two different algorithms, each designed for a different range of packet loss rates. Additionally, we recommend breaking up the whole picture into several rectangular slices and apply the two algorithms to the resulting slices, as suggested in Scenario 4.

For low packet loss rates of up to 5%, it is suggested and verified in the simulation results, that complete I-pictures should be transmitted frequently, rather than relying on the intra macroblock update mechanism of H.263+. The interval between two I-pictures should be chosen similarly to the average packet loss interval for low motion sequences and similar to one half of the typical packet loss interval otherwise. That is, if in average every 20^{th} packet get lost (corresponding to an observed packet loss rate of 5%), every 20^{th} sent picture should be an I-picture for low motion sequences and every 10^{th} sent picture should be an I-picture for high-motion sequences.

At higher packet loss rates of up to 20%, our suggestion is to use Video Redundancy Coding. This algorithm, which is completely based on inter picture prediction, yields only small improvements over the former algorithm in terms of bit rate and PSNR, but offers a significant

subjective quality gain and allows lower latency, simultaneously. Due to the scalability of VRC, the VRC parameters (number of threads and thread-length) can be chosen according to the error rate. In this work, we suggest to parameter sets as indicated in **Error! Reference source not found.**, which summarizes our recommended algorithms for Scenario 5, showing the two decision criteria (packet loss rate and intensity of motion) and the suggested algorithms.

Decision Criterion 1	Suggested Algorithm	Decision Criterion 2	Suggested Algorithm Parameters
<=5 %	Repetitive I-pictures	Heavy motion	I-picture Interval == mean packet loss interval
		Light motion	I-picture interval == ½ of mean packet loss interval
> 5%	Video Redundancy Coding	Packet loss <10%	VRC, 2 threads and 3 pictures per thread
		Packet loss >= 10%	VRC, 3 threads and 3 pictures per thread

Table 2: Scenario 5 algorithms for the different packet loss rates

5 Simulation results

To perform our simulation experiments, we used the University of British Columbia's H.263+ Reference Codec, which was extensively modified to support some of the error resilience optional modes. None of the coding efficiency oriented optional modes was used. Although hundreds of simulations have been performed, only a few results are discussed in the section, mostly because of space constraints. Other results, some of which can be found in a frequently updated technical report [29], were also used to arrive at the recommendations presented in the previous section.

In order to properly interpret the results, several objective/subjective quality measures should be used. It is well known that simple mathematical objective measures such as the PSNR do not correlate well with the subjective reproduced video quality. Even in still images, small artifacts, for example, seem to have a significant impact on the subjective quality of an image, although they do not decrease the PSNR significantly. There are numerous published methods (e.g., [30]) that introduce an objective quality scale for individual pictures, but such methods are still not widely accepted.

Quality measurement for video is even more difficult, especially at low bit rates. In fact, the experts of Q.15 of SG16 of the ITU-T, who are responsible for the standardization of coded representation of video, rely mostly on tedious subjective quality assessments. See the ITU-R recommendation 500-1 [31] for information about how subjective video quality testing should be performed.

Additional complexity is faced, if we have to consider events with a random nature, like packet losses. A packet loss, for example, which effects a P-picture immediately before a new I-picture is to be transmitted, will have virtually no visible impact. If, however, the packet loss effects data of that I-picture, a significant lack of visible quality will be observable for quite some time. For the above reasons, we rely mostly on subjective quality testing of reproduced video sequences, taking into account factors like skipped pictures and local artifacts in high-motion

areas. Note that our ratings are relative to the "best possible" reproduction quality obtained in the context of error-free transmission. This means, that even a 20kbit/s, 10 frames per second, encoded QCIF video sequence will be rated as "perfect" quality in terms of our standard, although its quality will be considered "bad" by most inexperienced users. The rating system outlined in Table 3 is used in our work. Finally, we also provide average PSNR over the luminance part of all the decoded video sequences. However, we do emphasize that PSNR values should be interpreted cautiously.

Rating	Definition
Perfect	The anchor for the specific simulation serious. No error-related artifacts visible.
Very good	Small Artifacts and single lost pictures cause minimal distortion.
Good	Larger artifacts and short "picture freezing" causes inexperienced observers to feel not completely comfortable with the quality.
Fair	Artifacts become larger, picture freezes up to some 100 ms
Poor	Artifacts like 'Ghost images' and picture freezes up to 500 ms do not justify a rating of "fair". It is, however, still possible to distinguish between artifacts and data.
Unusable	The quality is unacceptable. Artifacts and data are no more distinguishable. Objects can no more be identified. Video with picture freezing longer than 1 sec becomes impractical.

Table 3: Definition of the subjective quality rating

The simulation environment necessary to exactly reproduce our results is available by accessing our WWW site http://kbs.cs.tu-berlin.de. It consists of the codec-software, a packet-loss simulator, a bit-error simulator, and the error patterns we used. We found, that the same packet loss probabilities, but different error patterns, can lead to a difference of as much as 2 dB in PSNR ratings even for long sequences of a duration of 5 minutes we used. Thus using the same error patterns is essential for the exact reproduction of our results.

5.1 Simulation of Scenario 1

The algorithms suggested in Section 4.1 were verified during the standardization process, both by objective quality measurements (PSNR) and assessment of the subjective reproduced video quality. We do not feel it is necessary to reproduce such results again for this paper, but we instead refer the reader to the proposal/justification documents [32, 33, 34, 35], the combined core experiment specification [36] and the report of those core experiments [37].

5.2 Simulation of Scenario 2

The maximum specified error rate for ISDN is 10^{-8}, although many tests on ISDN-lines show an even smaller error probability [38]. At an error-rate of 10^{-8} and a bit rate of 110 kbit/s (which is commonly used in H.320-systems involving two B-channels), a bit error occurs approximately every 910 seconds. Therefore, H.263+'s error resilience modes do not really need to be used. However, since Annex H was originally designed for ISDN, we simulated the transmission of

video over ISDN with and without such annex. More specifically, we simulated the transmission of the bit stream output of a 15-minute segment of the CIF video sequence Paris coded at a fixed bit rate of 110 kbit/s and a fixed frame rate of 15 fps, with and without using Annex H. We have here used the CIF video format because it is the format typically used in practical ISDN systems. **Error! Reference source not found.** shows the results for four different cases: 1) No bit errors without Annex H, 2) no bit errors with Annex H, 3) 10^{-8} bit error rate without Annex H, and 4) 10^{-8} bit error rate with Annex H. Random bit errors were applied using the same error pattern for the two latter cases. As expected, no improvement in objective quality (PSNR) or subjective quality was observed. In fact, the application of Annex H requires a 4% increase in bit rate (leaving only 105.6 kbit/s for video), which yields a slight decrease in PSNR. Clearly, the use of Annex H cannot be justified, at least, based on our simulation results.

No bit errors		**10^{-8} bit error probability**	
Baseline	**Annex H**	**Baseline**	**Annex H**
Perfect, 28.86 dB	Perfect, 28.69 dB	Perfect, 28.86 dB	Perfect, 28.69 dB

Table 4: Subjective and PSNR ratings for BCH-FEC at ISDN error rates

5.3 Simulation of Scenario 3

For this scenario, the sequence Paris is coded at a fixed target bit rate of 20 kbit/s and a fixed frame rate of 10 fps, resulting in different picture quality levels (by using different quantizer values). The rate control method discussed in TMN8 [39] is employed in this simulation.

We use packets of a maximum of 1400 bytes of H.263+ payload data per packet. We also transmit a maximum of one coded picture in one packet, to meet both the RTP-payload specification's requirement and the practical needs of low-delay communication. All bit rates stated below include a packetization overhead of 40 bytes per packet, resulting in 400 bytes or 3200 bits of packetization data per second and thus limiting the effective video bit rate to 16800 bit/s.

Random packet losses of 0%, 3%, 5%, 10% and 20% were simulated, both with and without the use of the Reference Picture Selection mode (RPS). The two simulation experiments not using the RPS mode were performed by sending repetitive I-pictures as in Scenario 5. In case where the RPS mode is used, a reliable back channel is assumed. Results for typical round-trip times for the back channel messages of 100 ms and 300 ms are given in the second column of Table 5.

Packet loss rate	**PSNR RPS**	**PSNR Intra**	**PSNR no Intra**	**Subjective RPS**	**Subjective Intra**	**Subjective, no Intra**
0%	26.1 / 26.1	N / A	26.1	Perfect	Perfect	Perfect
3%	26.0 / 25.9	N / A	19.0	Very Good	Unusable	Fair
5%	25.9 / 25.9	N / A	19.0	Very Good	Unusable	Fair
10%	25.8 / 25.6	N / A	18.0	Very Good	Unusable	Fair
20%	25.5 / 25.3	N / A	14.4	Very Good	Unusable	Poor

Table 5: Subjective and PSNR ratings for Scenario 3's algorithms at different packet loss rates

Note that for the repetitive I-pictures case, no PSNRs were computed, because such would not be meaningful. The transmission of one I-picture takes up to 2 seconds at the given bit rate and for a reasonable chosen quantizer, leading to unacceptably high delays, especially at the higher packet loss rates. At a 10% packet loss rate, for example, the algorithm displays repetitively about one second of video, before the picture freezes for the duration of nearly two seconds. The subjective ratings shown in Table 4 reflect this situation.

5.4 Simulation of Scenario 4

For Scenario 4, we suggest that the picture be divided into independently decodable rectangular slices of a reasonable size (so that each coded slice fits into one packet). The Reference Picture Selection mode should be applied to each of the slices.

This simulation consists of three sets of experiments. The first set of experiments is intended to demonstrate the bandwidth increase incurred by the use of independent decodable rectangular slices. Therefore, we divide each picture of CIF size video sequences at a fixed frame rate of 15 fps into four QCIF size independently decodable rectangular slices (ISR), each covering one quadrant of the original CIF size picture. Then, we encode each of the slices using a fixed quantizer of 10. Table 6 shows the bit rate increase (in percentages) for each of the sequences Paris, Foreman and Coastguard along with the absolute bit rates. For each stream, those numbers can be compared to each other, because the fixed quantizer value was used leading to very similar quality. The bit rate increase is 8.7% for the low-motion sequence Paris, and somewhat higher (up to 13.2%) for the high motion sequences Foreman and Coastguard.

Stream	Bit rate CIF (kbit/s)	Bit rate ISR (kbit/s)	Bit rate increase
Paris	125.0	135.9	8.7%
Foreman	139.0	157.3	13.2%
Coastguard	247.3	278.9	12.8%

Table 6: Bit rate increase incurred by using Independent Decodable Rectangular Slices.

In the second set of experiments, the QCIF size video sequence Paris at a frame rate of 15 fps and a fixed quantizer 16 is coded, using the repetitive I-pictures (intra) algorithm and the back channel supported Reference Picture Selection (RPS) mode algorithm. For the repetitive I-pictures algorithm, an I-picture frequency as suggested in Section 4.4 is used. Table 7 shows the PSNR performance for both the intra and RPS mode algorithms for several packet loss rates and 100ms or 300ms round-trip delay of the reliable back channel messages. Clearly, the RPS mode algorithm achieves significantly lower bit rates at similar PSNR and subjective quality values.

Packet Loss	PSNR		Subjective Quality		Bit rate (kbit/s)	
	RPS	Intra	RPS	Intra	RPS	Intra
0%	27.1	26.8	Perfect	Perfect	43.8	35.7
3%	27.0 / 27.0	26.9	Perfect	Perfect	44.4 / 44.6	47.9
5%	26.9 / 26.9	26.8	Perfect	Perfect	44.9 / 45.2	56.0

10%	26.8 / 26.7	26.8	Perfect	Perfect	45.8 / 48.0	71.8
20%	26.7 / 26.4	26.9	Perfect./Very good	Perfect	47.8 / 52.8	118.3

Table 7: Advantage of RPS mode compared to a repetitive I-picture algorithm

In the third set of experiments, we also encode the QCIF size video sequence Paris at a frame rate of 15 fps, but now at a fixed bit rate achieved using the standard TMN8 rate control method. We compare those results (which rely on the usual intra macroblock update mechanisms of H.263+ as the only error resilience tool) the same coder employing the RPS mode algorithm. For a fair comparison, the TMN8 rate control method is used to select the same bit rates tested in the previous set of experiments. Moreover, the worse case of the two simulated round trip times for the back channel messages in the case of RPS (i.e., 300ms) is assumed. Table 8 shows the PSNR values achieved by the TMN8 coder and the coder employing the RPS mode algorithm for several packet loss rates. Again, the RPS mode algorithm offers a clear advantage, especially at the higher packet loss rates.

Packet Loss	Bit rate (kbit/s)	PSNR		Subjective Quality	
		RPS	TMN8	RPS	TMN8
0%	43.8	27.1	27.8	Very good	Perfect
3%	44.6	27.0	25.6	Very good	Good
5%	45.2	26.9	24.7	Very good	Good
10%	48.0	26.7	22.2	Very good	Fair
20%	52.8	26.4	20.3	Good	Poor

Table 8: Advantage of RPS compared to standard TMN8 coder

5.5 Simulation for scenario 5

For Scenario 5, we suggested two different algorithms, one for lower packet loss rates below 5% and the other for higher packet loss rates between 5% and 20%. Simulation experiments for both cases are presented next.

5.5.1 Simulation for the Packet Loss rates below 5%

To simulate this scenario for packet loss rates below 5%, we encoded the QCIF sequences Paris, Foreman, and Coastguard at a fixed bit rate of 110 kbit/s and a fixed frame rate of 15 fps. Due to rate control inaccuracies, the actual bit rates have values that are within +- 4% of the target bit rate. The packets (a maximum of 1400 bytes payload) underwent a random packet loss rate of 3% or 5%. Table 9 shows the PSNR and subjective performance for the encoded sequences at different packet loss rates of the TMN8 coder, the same TMN8 coder but with repetitive sending of I-pictures at packet loss intervals, the same TMN8 coder but with repetitive sending of I-pictures at ½ packet loss intervals, and the same TMN8 coder with Video Redundancy Coding (VRC) applied (with a fixed I-picture interval of 5 seconds). Although the PSNR does not show a significant difference, the subjective quality is much better when sending complete I-pictures

from time to time instead of relying completely on the update mechanism. Moreover, for high motion video sequences, a higher I-picture frequency should be selected, as recommended in Section 4.5.

Input Data, Packet loss rate	TMN8	TMN8, I-picture Interval == packet loss Interval	TMN8, I-picture Interval == ½ packet loss Interval	VRC 2-3, 5s I-picture Interval
Paris, 3%	27.7, good	30.0, good (Note 1)	29.8, good	30.2, good
Paris, 5%	25.2, fair	29.2, good	29.3, good	29.9, good
Foreman, 3%	29.3, good	31.2, good	31.2, good	30.4, good
Foreman, 5%	26.9, fair	29.6, good	30.8, good	30.0, good
Coastguard, 3%	27.4, good	28.9, good	29.8, good	28.8, good
Coastguard, 5%	25.5, fair	28.5, good	29.4, good	28.5, good

Table 9: Performance of standard TMN8 rate control versus repetitive I-picture

5.5.2 Simulation Experiments for Packet Loss Rates Above 5%

At higher packet loss rates, we employ Video Redundancy Coding. We next show some simulation results. Other results can be found in [8, 20]. Table 10 shows the results conducted using variable bit rate for a high quality QCIF-coding of the sequence Paris (with a fixed quantizer factor of 16). Packet loss rates of 10% and 20% were simulated by applying a random packet loss rate. Both 2 thread/ 3 frames per thread, and 3 thread, 3 frames per thread VRC parameters were simulated. As the table indicates, for the 10% case a 2-3 VRC is sufficient, whereas in the 20% case is more appropriate. For both non-zero packet loss rates, VRC performs much better than the standard TMN8 coder.

Packet loss rate	Coding scheme	Data rate (kbit/s)	PSNR	Subjective quality
0%	TMN8	31.9	27.1	Perfect
10%	TMN8	31.9	24.9	Poor
	VRC 2-3	43.2	26.3	Good
	VRC 3-3	50.1	26.5	Good
20%	TMN8	31.9	23.0	Poor
	VRC 2-3	43.2	25.1	Fair
	VRC 3-3	50.1	25.6	Good

Table 10: VRC performance compared to TMN8 at packet loss rates above 5%

6 Conclusion

In our paper, we present the error resilience oriented mechanisms of H.263+. It was shown that each of the error resilience oriented optional modes have specific advantages and applications. However, it is necessary to carefully decide for a given network/protocol/bit rate/error rate combination what modes should be applied.

By limiting our scope to widely deployed networks and protocol hierarchies, we are able to discuss those important cases in some detail. For each of the combinations, called scenarios throughout the paper, a combination of the optional modes is recommended that provide good performance.

Introducing error resilience to video coding clearly has several advantages. The proposed mode combinations for the described scenarios do offer acceptable performance especially in cases where latency is critical. Latency is a key parameter in interactive applications, like Videotelephone and Videoconferencing systems. We feel, that the design of high performance interactive video communication systems will rely in the future mostly on mechanisms similar to the ones described in our work.

7 List of Tables and Figures

7.1 Figures

7.2 Tables

8 References

[1] ITU-T Recommendation H.261: "Video codec for audiovisual services at p x 64 kbit/s", 1993

[2] ITU-T Recommendation H.320: "Narrow-band visual telephone systems and terminal equipment", 1997

[3] ITU-T Draft Recommendation H.263 V.2 (H.263+): "Video Coding for Low Bit rate Communication", Draft20, available from ftp://standard.pictel.com/q15-site

[4] G. Cote, B. Erol, M. Gallant, F. Kossentini: "H.263+: Video Coding for Low Bit Rates", accepted for publication in the special 1998 issue of the Transactions of Circuits and Systems for Video Technology

[5] Coding of moving pictures and associated audio for digital storage media up to about 1,5 Mbit/s, ISO/IEC International Standard 11172, 1992

[6] T. Nakai, Y. Tomita: "Core Experiments on Back-Channel Operation for H.263+" ITU-T SG15 contribution LBC 96-308, November 1996

[7] H. Kimata, Y. Tomita, H. Ibaraki, and T. Ichikawa "Concealment of Damaged Are for Mobile Video Communication", Proceedings AVSPN97, Aberdeen, U. K., 1997

[8] S. Wenger "Video Redundancy Coding in H.263+", Proceedings of AVSPN 97, Aberdeen, U. K., 1997

[9] Vinay Kumar: "Mbone: Interactive Multimedia on the Internet", New Ryders Publishing, ISBN 1562053973, 1996

[10] R. Schaphorst: "Videoconferencing and Videotelephony: Technology and Standards", Artech House, ISBN 0890068445, 1997

[11] See http://www.itu.ch for an overview

[12] See http://www.ietf.org for an overview

[13] ITU-T Recommendation H.324: "Terminal for low bit rate Multimedia Communication", 1996

[14] ITU-T Recommendation H.223 "Multiplexing Protocol for Low Bit rate Multimedia Communication", 1997

[15] See http://www.imtc.org for an overview

[16] ITU-T Recommendation H.221: "Frame structure for a 64 to 1920 kbit/s channel in audiovisual teleservices", 1997

[17] ITU-T Recommendation H.225.0: "Media stream packetization and synchronization on non-guaranteed quality of service LANs", 1996

[18] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. "RTP: A Transport Protocol for Real-time Applications." Proposed Standard. RFC 1889. January 1996

[19] M. Handley. "An Examination of Mbone Performance." UCL/ISI Research Report. January 1997.

[20] J. Ott, S. Wenger, G. Knorr: "Application of H.263+ Video Coding Modes in Lossy Packet Network Environments", submitted to Journal for Visual Communication, January 1998

[21] J.C. Bolot and A. Vega-García. "The Case for FEC-Based Error Control for Packet Audio in the Internet." To appear in ACM Multimedia Systems.

[22] M. Yajnik, J. Kurose, and D. Towsley. "Packet Loss Correlation in the Mbone Multicast Network." Proceedings of the IEEE Global Internet Conference. London. November 1996.

[23] M. Handley. "SAP: Session Announcement Protocol." Internet-Draft draft-ietf-mmusic-sap-00.txt. Work in progress. June 1996.

[24] M. Handley, H. Schulzrinne, and E. Schooler. "SIP: Session Initiation Protocol." Internet-Draft draft-ietf-mmusic-sip-04.txt. Work in progress. November 1997.

[25] C. Bormann, J. Ott, and C. Reichert. "Simple Conference Control Protocol." Internet Draft draft-ietf-mmusic-sccp-00.txt. Work in progress. June 1996.

[26] S. Casner, H. Schulzrinne, D. Kristol: "Frequently Asked Questions (FAQ) on the Multicast Backbone (MBONE)", http://www.mediadesign.co.at/newmedia/more/mbone-faq.html

[27] ITU-T recommendation H.242: "System for establishing communication between audiovisual terminals using digital channels up to 2 Mbit/s", 1996

[28] C. Bormann, L. Cline, G. Deisher, T. Gardos, C. Maciocco, D. Newell, J. Ott, G. Sullivan, S. Wenger, and C. Zhu. "RTP Payload Format for the 1998 Version of ITU-T Rec. H.263Video (H.263+)." Internet Draft draft-ietf-avt-rtp-h263-video-01.txt. Work in progress. January 1998.

[29] S. Wenger, G. Knorr: "Simulations of H.263+ Error Resilience Mechanisms", available from http://kbs.cs.tu-berlin.de, 1998

[30] M. Miyahara, K. Kotani, V. R. Akgazi. "Objective Picture Quality Scale (PQS) For Image Coding", submitted to IEEE Transactions on Communication

[31] ITU-R Recommendation 500.1: "Method for the Subjective Assessment of the Quality of Television Pictures", 1978

[32] ITU-T SG15 Contribution LBC-96-085: "Syntax Modifications for Reducing Video Delay", available from ftp://standard.pictel.com/q15-site

[33] ITU-T SG15 Contribution LBC-96-099: "A method for reducing delay in H.263 video coding"

[34] ITU-T SG15 Contribution LBC-96-132: "Proposed H.263 syntax changes for Error Resiliency and Packetization"

[35] ITU-T SG15 Contribution LBC-96-116: "GOB-Source-Formats (GSF) for H.263 (and an additional supervisory message for H.223)"

[36] ITU-T SG15 Contribution LBC-96-145: "Core Experiments for GOB Layer syntax"

[37] ITU-T SG15 Contribution LBC-96-245: "Results of the Core Experiments for GOB Layer syntax"

[38] ITU-T Recommendation I.430: "Basic user-network interface Layer 1 specification", 1996

[39] ITU-T working document Q15C15: "Video Codec Test model Near-Term, Version 8 (TMN 8)", available from ftp://standard.pictel.com/q15-site