H.263+: THE NEW ITU-T RECOMMENDATION FOR VIDEO CODING AT LOW BIT RATES

Thomas R. Gardos

Intel Corporation
5200 NE Elam Young Parkway
Hillsboro, Oregon 97124, USA

ABSTRACT

H.263+ is a revision to the 1996 version of ITU-T Recommendation H.263 that brings incremental improvements to compression performance, better support for packet-based networks, expanded support for video formats as well as other new functionality. All the new capabilities can be negotiated individually or disabled for backwards compatibility with H.263. In this paper, we review all the major new features of H.263+.

1. INTRODUCTION

H.263+[1] is a revision to the original 1996 version of ITU-T Recommendation H.263[2][3]. The original H.263 was developed for video compression at rates below 64 kilobits per second, and more specifically at rates below 28.8 kilobits per second. This was the first international standard for video compression which would permit video communications at such low rates.

In the original ITU-T work plan, the goal was to define a "near term" recommendation in 1996, followed by a "long term" recommendation several years afterwards. The near term recommendation is what is referred to as H.263. The long term recommendation (previously called H.26L) is currently scheduled for standardization in 1999 and may be a completely new compression algorithm. After H.263 was completed, it became apparent there were incremental changes that could be made to H.263 that visibly improved its compression performance. It was thus decided in 1996 that a revision to H.263 would be created which incorporated these incremental improvements. This would be H.263 "plus" several new features, hence the working name H.263+. H.263+ is scheduled to be finalized in February, 1998.

H.263+ contains approximately 12 new features that do not exist in H.263. These include new coding modes that improve compression efficiency, support for scalable bitstreams, several new features to support packet networks and error-prone environments, added functionality and support for a wider variety of video formats. An overview of these new features are presented here.

2. FUNDAMENTALS

H.263+ falls into the family of video coders commonly referred to as hybrid DPCM/DCT-based video compression algorithms. Other video coders in this family include H.261[5], MPEG-1, MPEG-2 and MPEG-4 video. These video coders typically employ 16×16 pixel block-based motion estimation and frame differencing to reduce temporal redundancy. A discrete cosine transform is applied to 8×8 pixel blocks of the resulting residual frame, and the transform coefficients are then quantized to reduce spatial redundancy. Lossless coding techniques are applied to the resulting symbols to further eliminate statistical redundancies. A frame that has been temporally predicted is referred to as an INTER coded or P-frame. The frame used as a basis of prediction, which is usually the decompressed version of the previous frame, is called the reference frame. When a frame is coded with no prediction whatsoever from a prior frame, it is referred to as an INTRA coded or I-frame. An INTRA coded frame is, for all intents and purposes, simply the result of still image compression applied to a frame of video. These components of video coders are described in great detail in numerous texts[4] and will not be discussed further here. The reader is assumed to be familiar with this general class of video coders. Here, we focus only on the new features of H.263+.

3. COMPRESSION IMPROVEMENTS

The following features contribute to improvements to compression performance in H.263+.

The compression performance of INTRA coded frames can generally be improved by exploiting the block-to-block correlation of DC and low-frequency horizontal and vertical transform coefficients. The *Advanced INTRA Coding mode* of H.263+ seeks to do just that. As shown in Figure 1, either the INTRA DC or first row or column of DCT coefficients are predicted from neighboring blocks. The selection of DC-only versus DC and AC coefficients, and the direction of prediction are indicated on a macroblock basis. When decoding, the prediction is performed after inverse quantization of the coefficient data. Moreover, two new zig-zag scan patterns are defined to better exploit the bias towards vertical and horizontal frequencies. Lastly, a new variable-length code (VLC) table is defined to better match the zero-run and amplitude statistics of INTRA coded frames. Previously, in H.263, the INTER frame VLC table was used for INTRA frames as well. This mode has been demonstrated to improve compression up to 40% on certain INTRA coded frames.

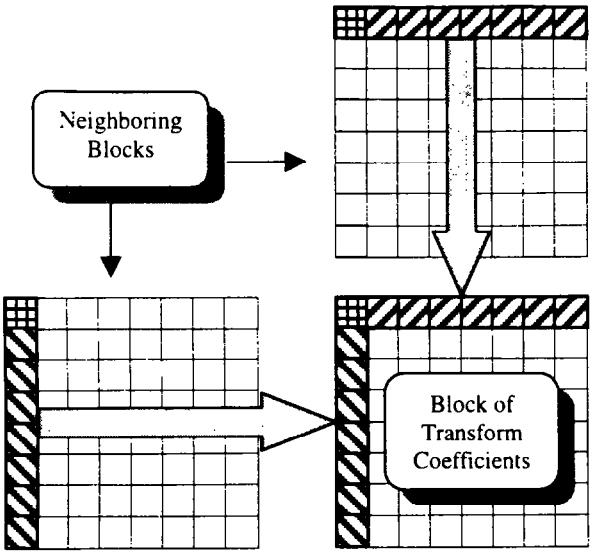

Figure 1 Coefficient prediction in Advanced INTRA Coding mode. Either just the DC coefficient is predicted or the entire first row or column of AC coefficients are predicted from the vertically or horizontally adjacent blocks respectively.

"Blockiness" is a common artifact in block-transform based video compression. It generally results from the lack of AC coefficients, representing high frequency detail, when reconstructing a transform block. The results are abrupt transitions at block edges, especially in regions that would normally contain smooth gradations. The *Deblocking Filter mode* reduces this effect by smoothing the transition at block edges. It operates on two pixels on each side of a horizontal or vertical block edge as shown in Figure 2. The pixel values A, B, C and D are replaced with new values according to a set of relations. The amount of smoothing is proportional to the quantization strength and inversely proportional to the pixel gradient at the block edge. The latter relation is so the filter is weakened at high amplitude edges since they are more likely to be actual edges in the scene content, rather than coding artifacts.

In the original H.263 Recommendation, there is a mode by which a bi-directionally predicted frame or B frame (as in MPEG-2) could be multiplexed with a subsequent P frame in the bitstream thus saving some of the overhead of having a separate frame in the bitstream. The motion vectors for the B frame would be interpolated from the P frame motion vectors and a small offset to the interpolated vectors could also be transmitted in the bitstream. Although in most cases, a P-B pair of frames would be more efficient than a pair of P frames, there were situations were this arrangement would perform worse. In H.263+, an improvement to the original mode allows a complete motion vector to be transmitted for the B blocks, thus a P/B frame pair can do no worse than a pair of P frames. This is referred to as *Improved PB-frames mode.*

H.263+ provides the means by which the reference picture can be resized, translated or generally warped before being used as prediction for the current frame. This is referred to as *Reference Picture Resampling (RPR) mode.* Several of the capabilities in

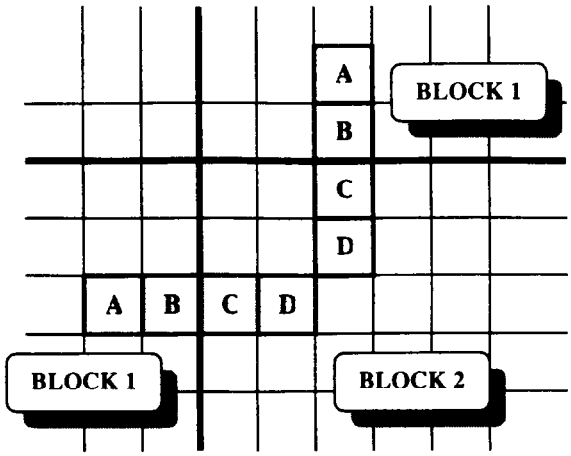

Figure 2 Illustration of the Deblocking filter. The filter is selectively applied to two pixels on either sides of vertical and horizontal block-edges according to a set of relations.

RPR mode can positively impact compression performance. For example, it is possible to indicate a global motion parameter. Say for instance that the video being compressed is undergoing a linear horizontal pan. Chances are most of the block-based motion vectors will represent the pan fairly accurately, however small deviations will result in excess bits when the motion vectors are differentially coded. A global motion parameter may be able to describe the global motion, thus allowing most of the block based motion vectors to be zero. Another useful sub-mode of RPR can be applied to frame size changes. In the original H.263 Recommendation, if the size of frames into the encoder changed (say from 176×144 to 352×288), it would be necessary to insert an INTRA coded frame since there was no valid reference frame. At very low data rates, INTRA coded frames are very expensive to encode as they can be from two to ten times the size of an INTER coded frame. The effect would be a long pause in video display while a large INTRA coded frame was being transmitted. With reference picture resampling, however, the reference picture could be resized to the new format and coding can continue with INTER coded frames. Another use for reference picture resampling is as an extreme bit rate control measure. Say for example that 352×288 pixel video is being compressed to very low data rates. In certain situations, usually under high motion, a frame cannot be compressed small enough even with the highest level of quantization. Typically this causes "jerkiness" or frozen frames in the video. With RPR, an encoder could dynamically switch to quarter sized video until the high motion subsides, thus avoiding "jerky" video. The display would continue to be full size, however the video would have to be interpolated back up to full size prior to display. In addition to the sub-modes just described, RPR permits general warping via arbitrary displacements of the four corners of the current frame of video with respect to the reference.

The *Reduced-Resolution Update (RRU) mode* is a slightly more esoteric technique for employing the extreme rate control measure described above. In reference picture resampling, the reference picture would be reduced to the size of the current picture, the compressed frame would be transmitted at the smaller frame size, and the result would be decoded and interpolated back up to the original size prior to display. The visual effect of

this operation would be that any high frequency detail that would have accumulated, say in the background of the video, would suddenly disappear when the reference picture was resampled. RRU mode addresses this problem by keeping the reference and current frames at their original sizes and down-sampling the residual frame after motion estimation/frame differencing to produce a valid quarter size INTER coded frame. In order to do so, however, macroblocks and blocks are redefined to be 32×32 pixels and 16×16 pixels respectively so that when the residual is downsampled, they return to their original sizes of 16×16 and 8×8 pixels. With RRU mode, areas of high motion can be accomodated without the side effect of detail loss. The tradeoff is the added complexity of implementing this mode of operation.

Typically, INTER coded frames at low data rates contain many low amplitude quantized coefficients with long runs of zeros between them. At high data rates (or finer quantization) the statistics change so that there are more large amplitude quantized coefficients and fewer long runs of zeros. This more closely resembles the statistics of INTRA coded frames. As part of the Advanced INTRA Coding mode, a new variable-length code (VLC) table was defined for just this type of statistic. Hence the compression performance for INTER coded frames at high data rates can be improved by using the VLC tables for Advanced INTRA. This is referred to as *Alternate INTER VLC mode* in the H.263+ text.

Among the new features of H.263+, one of several which corrects design inefficiencies of the original H.263 Recommendation is *Modified Quantization mode*. This mode has four key elements: indication for larger quantizer changes from macroblock-to-macroblock to better react to rate control requirements; the ability to use a finer chrominance quantizer to better preserve chrominance fidelity; capability to support the entire range of quantized coefficient values rather than having to clip values greater than about 128; and explicitly restricting the representation of quantized transform coefficients to those that can reasonable occur.

Another modification to the original H.263 Recommendation concerns motion vector range. Whereas before, the motion vector range was for the most part [-16,15.5]. When H.263+ mode is invoked, the range is generally larger and depends on the frame size as shown in Table 1.

Another modification to the original H.263 Recommendation is the addition of a rounding term to the equation for half-pel interpolation. Without it, there is a positive bias in the half-pel interpolation which can most notably be seen as a pink color drift in facial flesh tones. The rounding term toggles from frame to frame to eliminate this rounding bias reducing the artifact noticeably.

Finally H.263+ supports a wider variety of input video formats than H.263. In addition to five standard sizes, arbitrary frame sizes, in multiples of 4, from (32x32) to (2048x1152) can be supported, as well as other pixel aspect ratios besides 12:11, and other picture clock frequencies besides 29.97 Hz.

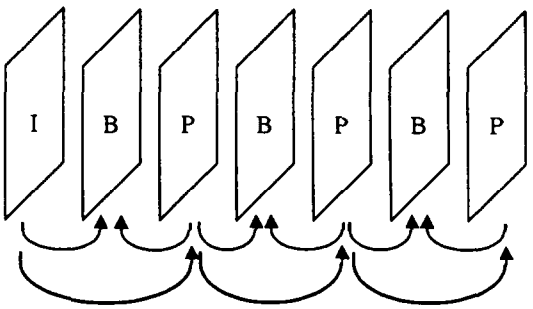

Figure 3 Bi-directionally predicted (B) frames.

4. SUPPORT FOR PACKET NETWORKS

On packet-based networks, the video bit stream must be fragmented and transmitted via numerous paths, then reassembled at the receiver. This has a number of implications on the video: there is usually no guarantee of arrival of packets, no guarantee that packets will arrive in order, nor even that the transmission times would be the same for each packet. Packet loss rates over networks such as the Internet can be 10% or higher. H.263+ has several new features that improve performance under these conditions. The first is the layered bit stream capability that can be used to prioritize data that, when selectively deleted, can reduce network congestion. Second is improved bitstream fragmentation capability. Third is a more robust mechanism for selecting reference pictures for prediction, and finally, the capability to define independent sub-pictures of a frame. These are discussed next in the context of packet-based networks.

Table 1 Motion vector ranges in H.263+.

Frame Sizes Up to...	Motion Vector Range
352×288	[-32,31.5]
704×576	[-64, 63.5]
1408×1152	[-128, 127.5]
Widths up to 2048	Hor. Range [-256, 255.5]

4.1 Temporal, SNR and Spatial Scalability

Most video codecs generate a single bit stream of compressed video that represents a particular quality level. In scalable video there are several complementary bit streams associated with a single video. One of the streams is called a base layer and represents the baseline quality of the video sequence. The other layers represent enhancements to the base layer. If a video application decoded an enhancement layer in addition to decoding the base layer, a better quality video could be reproduced.

H.263+ video enhancement layers belong to one of three categories: temporal, SNR or spatial enhancement. Temporal enhancement is the process by which the frame rate can be increased over the base layer. This is accomplished via disposable bi-directionally predicted frames, or B-frames for short. The bi-directionally predicted nature of B-frames are illustrated in Figure 3. Predicting from prior and subsequent frames usually improves compression performance. The

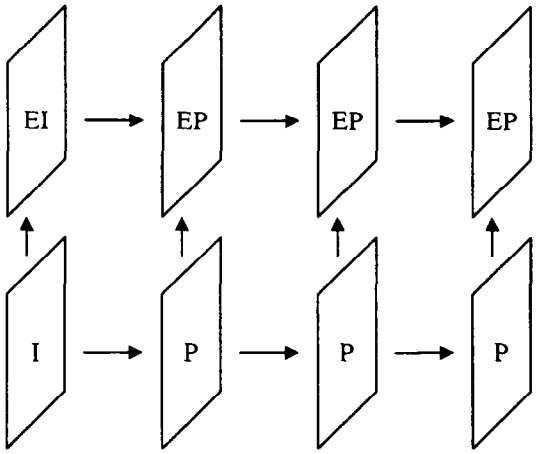

Figure 4 SNR enhancement layer

important aspect to note is that B-frames are not used as the basis for prediction for any other frames so they are truly disposable. They are not required to decode the I and P-frames, however, when added, they raise the frame rate of the video sequence. Hence the name temporal enhancement layer.

The SNR enhancement is a refinement to the coded base layer frames (see Figure 4). When a video frame is compressed, the decoded version is not an exact replica of the original frame. This is because H.263+ is a lossy compressor - it selectively throws away information in order to improve compression performance without excessive degradation of visual quality. Because the decoded frame is not an exact replica of the original, there are non-zero difference values when the decoded frame is subtracted from the original. This is sometimes called the residual data - this is what is left over after a frame has been compressed. In SNR scalability, this residual is coded separately and sent as a separate layer in the bit stream (the EI and EP frames in Figure 4). If there is enough bandwidth for a decoder to receive the SNR enhancement layer, the visual quality can usually be improved over the base layer. Besides the prediction from the base layer, the H.263+ encoder has the choice of including prediction from the previous frame in the SNR enhancement layer as well. This is a modified form of a bi-directionally predicted frame (called EP frame in this case). When an SNR enhancement frame is only predicted from the base layer and not a previous frame, it is called an EI frame.

Spatial enhancement is closely related to SNR enhancement. The only difference is that the enhancement layer is twice the size vertically and horizontally, from the base layer. In this case, the input video is first down-sized both vertically and horizontally prior to encoding as the base layer. Then, the decompressed base layer frames are interpolated back up before being used as prediction for the spatial enhancement layer frames.

More than one enhancement layer can be used to create multiple layers in a bit stream. Moreover, different enhancement types can be combined to provide a very flexible layered bit stream architecture.

4.2 Packetization and Error Resiliency

Part of the process of transporting video bitstreams over packet-based networks is fragmentation of the bitstream before being converted into a packet payload. H.263+ improves support for this fragmentation operation through the use of arbitrary resynchronization markers. The resynchronization markers are provided by slice headers in a sub-mode of the *Slice Structured mode*. Slice headers can be inserted at any macroblock boundary.

When packet loss becomes substantial, it is possible to define sub-pictures within each picture which do not depend on any information outside the sub-picture boundaries within the current frame or in any referenced picture. This way, errors due to packet loss can be kept from propagating to other regions. This is referred to as Independent Segment Decoding mode.

H.263+ also provides the ability to indicate (either positively or negatively) that a prior frame is suitable to be used as a reference for compression. This helps ensure that a sender does not use a reference frame that has been corrupted in the decoder. This is referred to as Reference Picture Selection.

5. SUMMARY

We presented a brief overview of the major new features of H.263+, the revision of ITU-T Recommendation H.263 due to be finalized in February, 1998. H.263+ brings numerous new optional features and capabilities to H.263. These include compression efficiency improvements such as improved INTRA coded frame compression, a deblocking filter, modified quantization and alternate variable length code tables. Furthermore, the reference picture resampling and reduced resolution update modes improve performance under certain situations. Besides compression improvements, there is support for scalable bitstreams, network packetization support, the ability to define independent sub-pictures for error containment, and the capability to acknowledge and identify error free reference frames. Lastly, H.263+ supports a wider variety of input formats than H.263.

6. REFERENCES

[1] *"H.263+", Video Coding for Low Bit Rate Communication, ITU-T Draft Recommendation H.263 Version 2*, International Telecommunication Union, September 26, 1997.

[2] *Video Coding for Low Bit Rate Communication, ITU-T Recommendation H.263,* International Telecommunication Union (March, 1996).

[3] Rijkse, K., *"H.263: Video coding for low-bit-rate communication,"* IEEE T-COM, Vol. 34, Issue 12, pp. 42—45, Dec. 1996.

[4] Tekalp, M., Digital Video Processing, Prentice-Hall, Inc., 1995

[5] *Video codec for audiovisual services at px64 kbit/s, ITU-T Recommendation H.261,* International Telecommunication Union (March, 1993)