ISO/IEC JTC1/SC29/WG11
MPEG92/
January 1992

CCITT SGXV
Working Party XV/1
Experts Group for ATM Video Coding

Document AVC-181
January 1992

SOURCE: Australian "Universal Video Codec" project (Telecom Australia Research Laboratories, Monash University, Siemens Ltd, University of New South Wales)

TITLE: Efficient Flexible Layered Coding

PURPOSE: Information

Abstract

Both the CCITT SGXV Experts Group for ATM Video Coding and the ISO/IEC Moving Picture Experts Group are developing coding schemes which should be suitable for a wide range of applications. To meet the diverse requirements which arise from this objective a flexible layering architecture is required. Layered coders are thought to be less efficient than non-layered coders. In this document the sources of inefficiency which arise in layered coders, and methods which can be used to reduce inefficiency, are described. Experimental results are presented to show that layered coders can be quite efficient in comparison with non-layered coders. The general principles should be useful in the design of efficient layered coders.

1. Introduction

The CCITT SGXV Experts Group for ATM Video Coding and the ISO/IEC JTC1/SC29 Moving Picture Experts Group (MPEG) are developing coding schemes which must be suitable for a wide range of applications. Typical applications are distinguished by their different quality requirements, different coding rates and different terminal complexity requirements. To provide a suitably large user community, in an environment where terminals with a range of capabilities exist, also requires coding schemes with interworking capabilities. Resolution/Bit-stream Scalability, i.e. the availability of low resolution version of the signal from a subset of the bit-stream, has been recognised as a fundamental requirement of the MPEG Phase 2 algorithm. Two examples where interworking/scalability will be essential are HDTV/TV distribution, where service providers will be interested in reaching a wide audience, and video-phone/video-conferencing, where users will need to communicate transparently with a range of terminal types, including those which have already been developed for ISDN. There may also be applications where interworking is not required.

To meet these diverse requirements a very flexible coding architecture is required and a single coding algorithm, such as CCITT Rec. H.261, or MPEG Phase 1, is unlikely to be suitable. Providing an inflexible coding standard will certainly impede both the development of a wide range of video services and the extensive use of video based services.

2. Flexible Layering

In previous contributions to the CCITT SGXV Experts Group for ATM Video Coding, Australia has proposed *flexible layering* as a general architecture which is suitable [AVC-35, AVC-73,AVC-135]. Flexible layering is an overall architecture, with many possible implementations using certain subsets of the full architecture to provide the required functionality and quality for different applications at different complexities. The general concept of flexible layering, which is an extension of the layered coding concept, is illustrated in figure AVC-181/1.

Flexible layering removes the constraint, usually implicit in layered coding, that coders must code all layers from the lowest resolution up to the desired service resolution. Rather, terminals only code the layers which are required, according to the range of other terminals with which interworking is desired, to provide appropriate cell loss concealment, or to provide flexibility in the trade-off between cost, complexity and performance. Examples are shown in figure AVC-181/1. The layers used during a connection could be configured at call establishment and could be dropped or added during a call, depending on the service requirements. Users and service providers can choose a layer configuration, which fits within the general architecture, and meets the requirements of their applications. Single layer coders are included as one special case. Many of the architectures put forward in contributions to the ATM Experts Group and to MPEG implement flexible layering [Proposals 8,14,18 and 38]. Proposal 23 also implements a very flexible coding system.

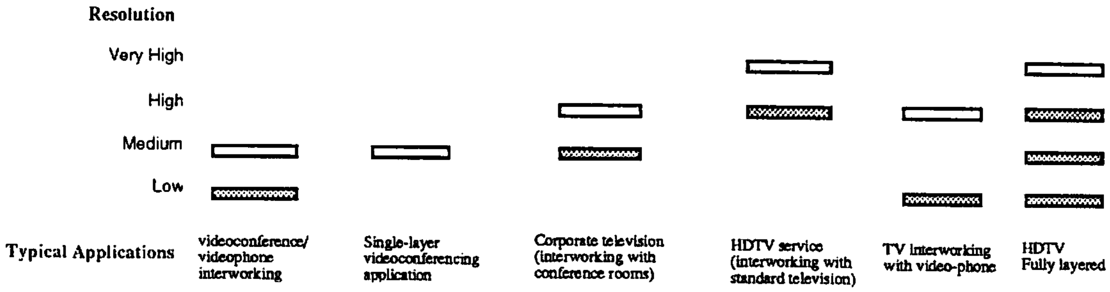

Resolution
Very High
High
Medium
Low
Typical Applications
videoconference/ videophone interworking
Single-layer videoconferencing application
Corporate television (interworking with conference rooms)
HDTV service (interworking with standard television)
TV interworking with video-phone
HDTV Fully layered

Figure AVC-181/1. A range of possible layered coders which fit within the flexible layered coding architecture. The shaded boxes represent optional layers, which can be switched off if not required.

3. Layered Coding Efficiency

It is clear that a flexible architecture is important for dealing with a range of applications. However, finding an implementation of flexible layering which provides the required functionality and also codes efficiently is important. It is well known that straight-forward layered coders are less efficient than conventional single layer coders. Our experiments have been aimed at identifying sources of inefficiency and ways of improving the efficiency of layered coding.

In our investigations we have identified the following factors which have an impact on layered coding efficiency:

Decomposition

A layered coder must decompose a signal into multiple layers. Using a critically sampled decomposition process aids coding, however, to provide flexible layering critical sampling is not possible. Different layers may have different quality requirements and, in the extreme case, some layers may be completely switched off. Therefore layer signals must include *both* information which represent additional resolution (i.e. higher spatial frequencies) and information which represents refinement quantization of the signal in lower layers. The impact of non-critical sampling can be minimised by embedding the decomposition process within transform coding. That is, by forming layer signals from a partitioning of the transform coefficients, rather than as a separate filtering operation. A number of the current MPEG proposals use this technique [Proposals 14 and 18]. In the upper layers this allows separation of the information associated with increased spatial resolution and information associated with refinement quantisation (critical sampling results when the refinement signal is zero). This separation is not possible in conventional pyramid coding systems.

Refinement Signal Coding

Refinement coding of lower layer signals is a potential source of coding inefficiency which will not be present in non-layered coders. Some layered coding schemes code refinement information in intra-frame mode only. This is inefficient since in stationary regions the fine quantization in the upper layers will continually code the inter-layer prediction error, even though the same correction was applied in previous frames. Allowing the refinement quantization to be coded in inter-frame mode avoids this problem.

Multiple Coding Modes

The availability of different coding modes for low frequency, high frequency, and refinement information improves the efficiency of layered coders in comparison with non-layered coders.

Motion compensation of layer signals

Decomposition involves filtering and decimation. Decimation will introduce aliasing distortion (the amount depends on filter design). A linear system which introduces aliasing is shift-variant and motion compensation cannot operate as effectively, since a shifted input no longer produces a shifted output. Note that layering by transform coefficient partitioning necessarily introduces aliasing. Shift-variance has two main consequences:

1. A single pixel shift in the full resolution image produces sub-pixel shifts in lower layer images. To compensate effectively for this motion requires sub-pixel accuracy in the lower layers.
2. Shift-varying distortion (aliasing), which can arise in all layer signals, including the upper most layer signal, cannot always be motion-compensated. One way of reducing its impact is by predicting the aliasing components. When decomposing using transform coefficient partitioning this can be achieved by passing blocks of the previous decoded full resolution frame through the decomposition process before using them in prediction.

In addition, the block size used for motion compensation should be appropriately scaled in all layers so that the same region in the scene is covered by a single block. This implies that exactly the same number of motion vectors are used in all layers.

Side Information (Start Codes, block addresses and types, motion vectors, etc)

Side information can become a significant overhead when coders operate at very low bit rates. Some of this information, such as the Start Codes and Macroblock Addresses is essential for synchronization and cannot be reduced. However, it is possible to use the information from lower layers to reduce the information required at upper layers. For example a hierarchical motion vector representation can be used so that the upper layers only code incremental motion vectors.

The architecture which has resulted from our experiments is described in appendix A. Currently it is a two layer architecture which provides CIF/QCIF resolution signals or, in single layer mode, a CIF resolution signal. The general principles which are highlighted should be useful in the design of layered coders operating at a variety of signal resolutions. This architecture is very similar to the Multi-purpose coding scheme (MUPCOS) [Proposal 18].

4. Summary of Experimental Results

Detailed results illustrating the comparative performance of our experimental coder in single layer and two layer coding modes are given in Appendix B. The results also show the comparative performance with a single layer RM8 coder. All coders operate in open loop Variable Bit Rate mode. In summary these results show:

1. Two-layer and one-layer coders, without loop filtering, provide approximately similar performance on a number of sequences (Table Tennis, Tempete, Trevor), over a range of bit-rates (500 Kbit/sec-3Mbit/sec).
2 Results for the sequence Flower Garden at high rates illustrate that layered coders become less efficient when a sequence contains moving high-frequency components. Under these conditions the layered scheme demonstrated is slightly better than simulcast.
3. The results for the sequence Trevor at low bit-rates (<400 Kbit/sec) indicate that when overhead information dominates the layered coder also becomes less efficient than a single layer coder, though it is still superior to simulcast.
4. When loop filtering is included in the single layer coder, coding performance improves for sequences containing predominantly low frequencies (e.g. Trevor). For sequences with moving high frequency components (e.g. Flower Garden) the results with loop filtering do not improve significantly and are worse at high bit-rates than the results without a loop

filter.

5. When compared with a two layer coder the single layer coder with loop filtering is generally superior at lower coding rates (<1.5 Mbit/sec). Note that our single layer coder is always better than the RM8 implementation of CCITT Rec. H.261.

Further work is required to determine appropriate loop filter designs for layered coders.

We note that, although the experimental results demonstrate a two-layer coding system using CIF and QCIF, the general principles which have been used to improve coding efficiency certainly apply in general to layered coders and can be used in the design of multi-layer coders operating at CCIR Rec. 601 resolution, or higher. Experiments at these higher resolutions are currently underway.

5. Conclusion

A coding architecture designed for video service coding must be flexible if it is to be of widespread use. Interworking/Scalability is one important requirement. Flexible layered coding is a means of achieving this objective. Our investigation have shown that layered coders can be quite efficient in comparison with single layer coders. By careful design many potential sources of inefficiency can be minimised and efficient flexible layered coders developed.

It is hoped that the information provided here will be of assistance in refining the CCITT/MPEG Test Model to allow it to satisfy the agreed requirements for scalability and maximum commonality across many applications, which still maintaining high coding efficiency.

Appendix A Experimental Coding Architecture

The coding scheme shown in figure AVC-181/2 can be described as a combined subband/pyramid coder. It uses DCTs to decompose a CIF-size image into low and high subbands. The low band signal, LL, at layer 0, which consists of the lower 4x4 coefficients of a DCT, can be used to reconstruct a QCIF resolution image. This low resolution image is coded by a VBR H.261-like coder, with an arbitrary but normally coarse quantizer Q0. The refinement signal at layer 1 is also coded by a VBR H.261-like coder with a finer quantizer Q1. The bit stream generated at layer 1 consists of the refinement quantization of the residue low band signal, ΔLL, as well as the high band signal HB (the upper 48 coefficients from an 8x8 DCT). This 2-layer structure can be extended into a multi-layer subband/pyramid structure if required. This structure is very similar to that put forward in proposal 18 (MUPCOS).

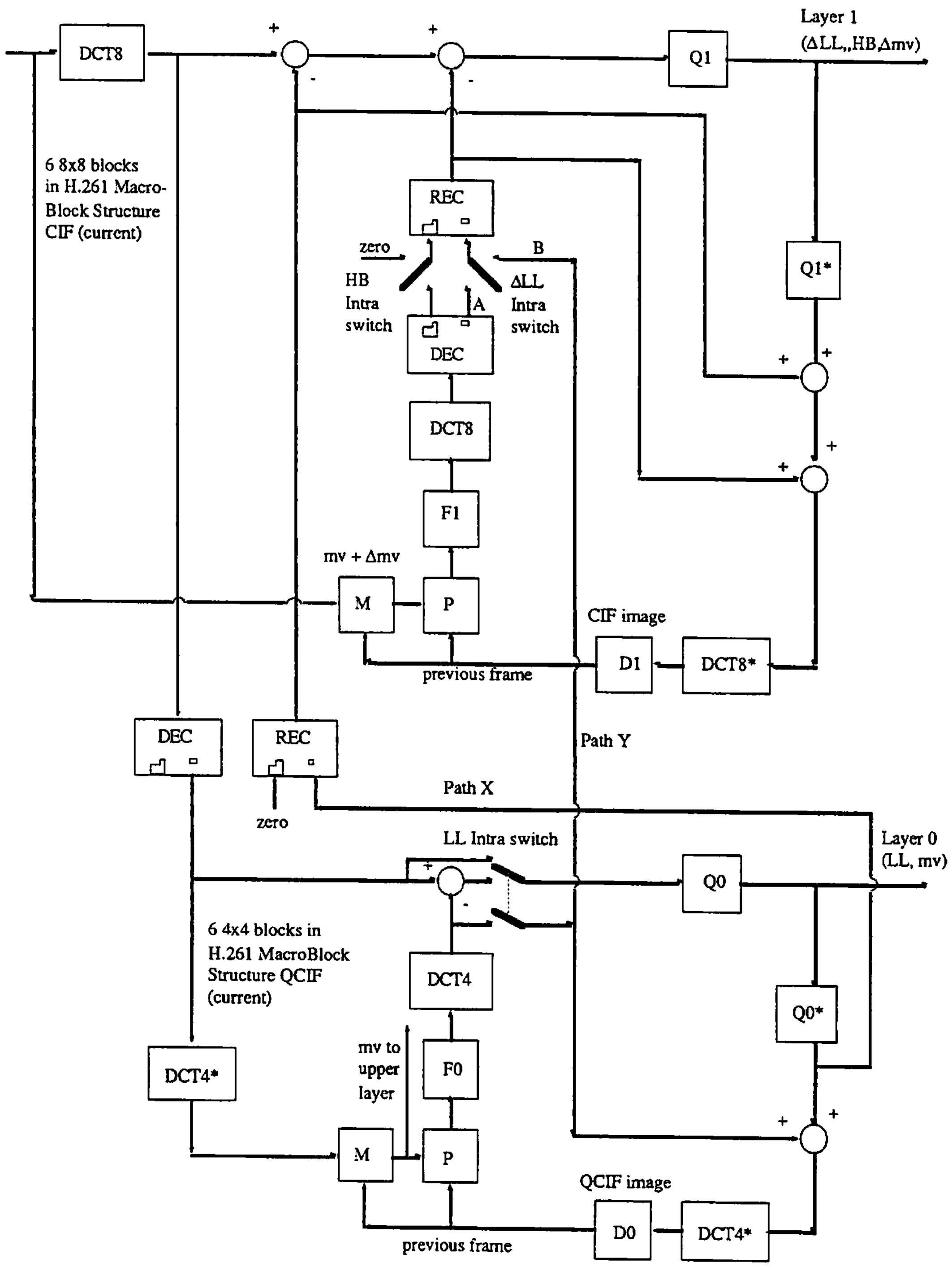
The coding architecture has two possible signal paths for inter-layer coupling. Note that the flexible layering requirement constrains the coupling to be one-way, i.e., from the lower layer to the upper layer. The first path labelled X in figure AVC-181/2 subtracts the decoded layer 0 coefficients from the low band DCT coefficients of the CIF resolution image. The second path, labelled Y in figure AVC-181/2, consists of the layer 0 predicted low band coefficients from the previous frame, and is in-place only when the ΔLL Intra switch is at position B. The overall layer structure has three switched decisions. In the lower layer, the coder can operate in intra-frame or inter-frame/inter-frame MC, as determined by the switch labelled LL intra. In the upper layer the HB coefficients can likewise be coded in intra-frame or inter-frame/inter-frame MC. The mode is determined by the switch labelled HB Intra. The coding mode used for the refinement quantization of the low band coefficients in the upper layer depends on both the LL Intra and ΔLL Intra switch positions. There are three modes used:

1. When the lower layer is in intra mode, the upper layer is constrained to operate in intra mode. The ΔLL signal is simply a refinement quantization of the lower 4x4 coefficients.
2. When the lower layer is in inter mode, and the ΔLL switch is at position A, the ΔLL signal consists of an inter-layer prediction of the prediction error.
3. When the lower layer is in inter mode, and the ΔLL switch is at position B, then ΔLL signal is again a refinement quantization of the lower 4x4 coefficients.

Note that mode 2 is very efficient for stationary regions, since the refinement quantization will be effectively coded in inter-mode, reducing the overall coding overhead. The greater adaptability afforded by the multiple decisions available in the layered structure is one reason for its high bit rate efficiency.

The layered architecture is also amenable to the implementation of block-based hierarchical motion estimation. The motion compensation block-size in layer 1 is 16x16 while the block-size in layer 0 is 8x8. The motion vector at layer 0 is obtained by full search block matching within a +/- 7 pixel search area using single pixel accuracy and a further +/- 1 pixel search, using interpolated image points, to obtain a sub-pixel accuracy motion vector. This motion vector is passed on to layer 1 as the initial vector. A ± 1 pixel search around this vector is then performed to find the layer 1 motion vector. The reference images used for block matching reside in the two frame stores D0 and D1. These two images are the decoded full band images of QCIF and CIF resolutions respectively.

The layer structure is flexible. Layer 0 can be dropped and the top layer reduces to a single layer H.261-like coder. In our current implementation single layer coding is indicated to the decoder with a flag in the bit-stream. In single layer mode the paths X and Y are no longer necessary and are set to zero. The two switches in the upper layer (HB intra and ΔLL intra) are "tied" together to give a single intra/inter-MC switch. In single layer mode the upper layer implements a full motion vector search. The single layer coder is compatible with H.261 at the macroblock level, however, since a SLICE structure is used instead of a GOB structure, a H.261 decoder cannot decode the bit stream generated by the coder.



DCT4: 4x4 DCT Transform
DCT8: 8x8 DCT Transform
DEC: Decomposition
Qn: Quantization

DCT4*: 4x4 Inverse DCT
DCT8*: 8x8 Inverse DCT
REC: Recombination
Qn*: Reconstruction

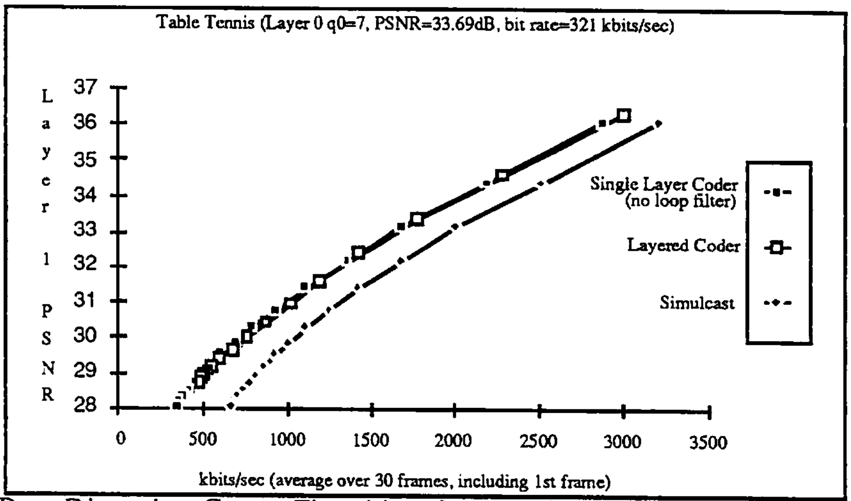
Dn: Frame Delay
M: Motion Estimation
P: Predictor
Fn: Loop Filter

Figure AVC-181/2. Block diagram of the Flexible Layered Coder.

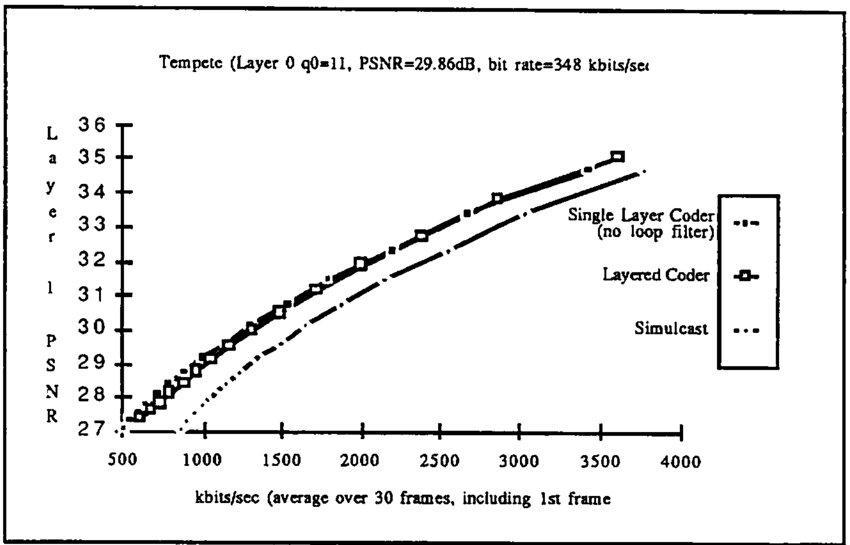
The coder currently uses the H.261 VLC code table. This code table works reasonably well but is by no means optimum. Because of the different block sizes, the H.261 code table contains events that will never occur at layer 0. Work is under way to design separate code tables for LL, ΔLL and HB components. The code tables would be different as the subband components all have different statistical characteristics.

Appendix B. Experimental Results

The layered coder described in Appendix A has been simulated and its performance investigated both in one-layer and two-layer modes. The performance has also been compared with CCITT Rec. H.261 (RM8). The rate distortion curves shown were obtained by varying the Layer 1 quantization step size while keeping the Layer 0 quantization step size fixed. All measurements were averaged over thirty frames of each sequence including the first frame.

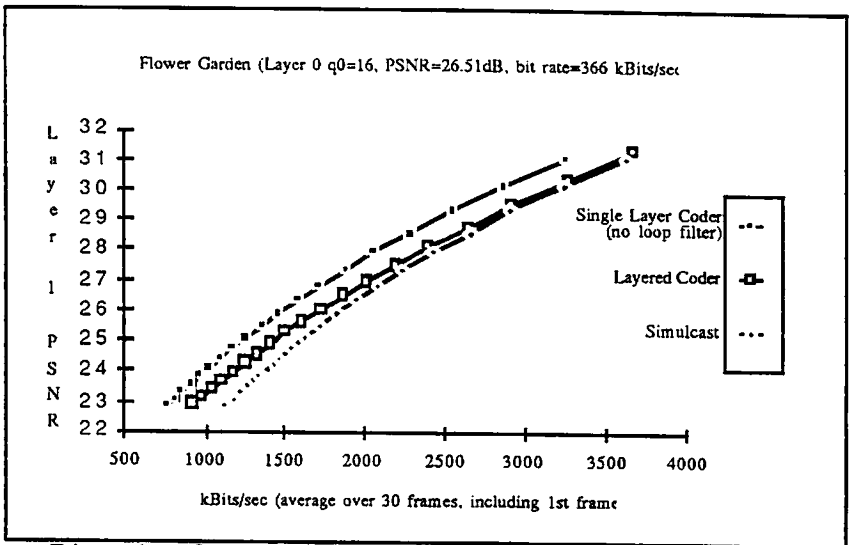


Graph 1(a) Rate Distortion Curve. First thirty frames of Table Tennis. No loop filter.

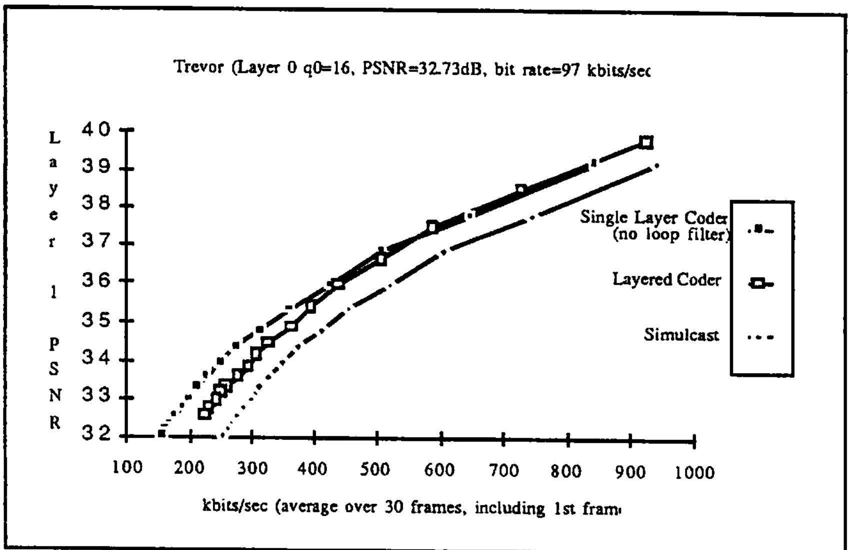


Graph 1(b) Rate Distortion Curve. First thirty frames of Tempete. No loop filter.

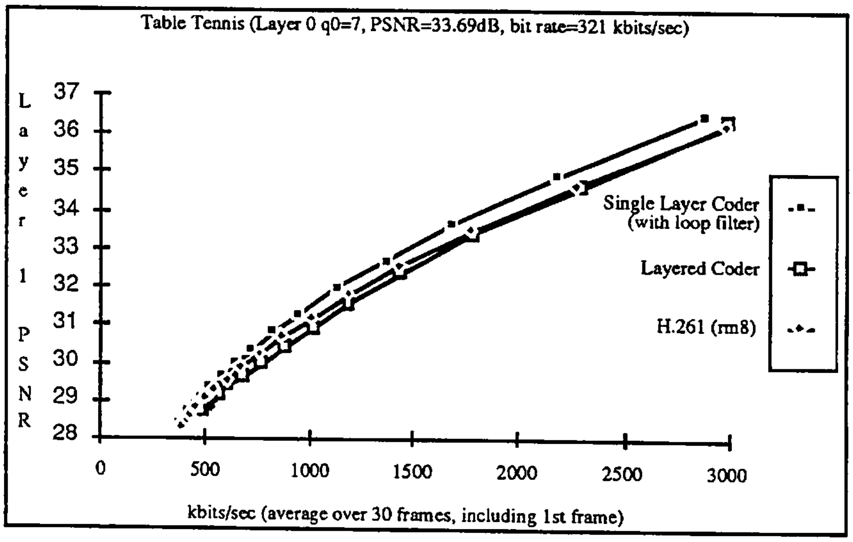
The first set of graphs 1(a) - 1(d) show that the layered coder compares favourably with the single layer coder when both have no loop filter. The second set of graphs, 2(a) - 2(d), show that at low bit-rates loop filtering improves the performance of the coder in single layer mode. We are currently investigating appropriate loop filter designs for the layered coder. The second set of graphs also show that the single layer coder is comparable with, and in some instances better than, the CCITT Rec. H.261 (RM8) codec.



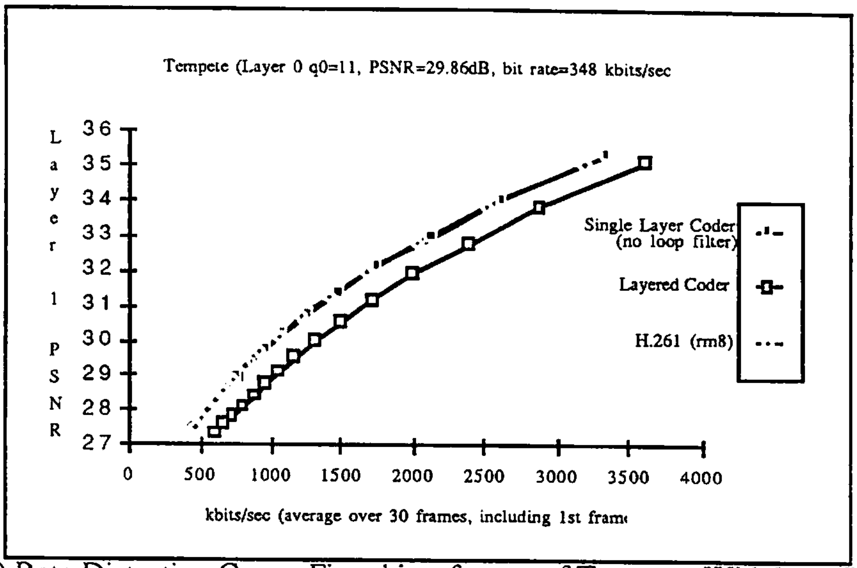
Graph 1(c) Rate Distortion Curve. First thirty frames of Flower Garden. No loop filter.



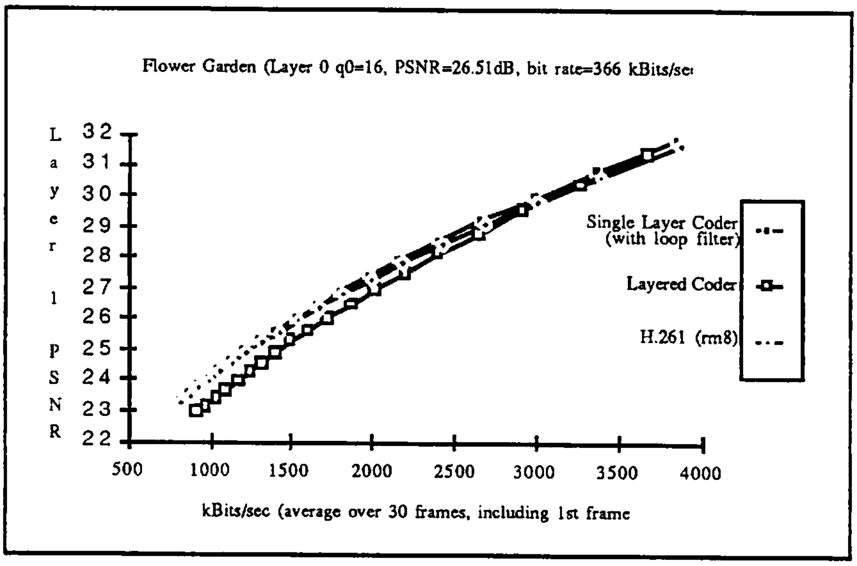
Graph 1(d) Rate Distortion Curve. First thirty frames of Trevor. No loop filter.



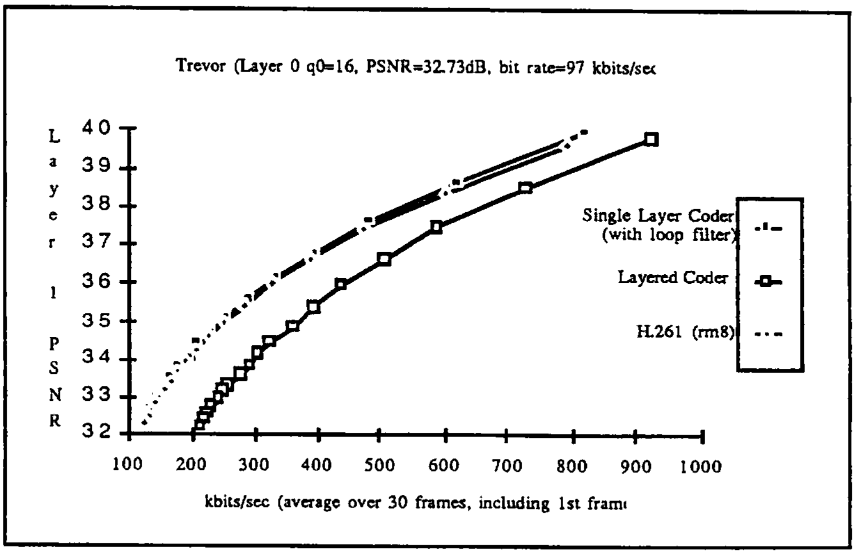
Graph 2(a) Rate Distortion Curve. First thirty frames of Table Tennis. With loop filter.



Graph 2(b) Rate Distortion Curve. First thirty frames of Tempete. With loop filter.



Graph 2(c) Rate Distortion Curve. First thirty frames of Flowers. With loop filter



Graph 2(d) Rate Distortion Curve. First thirty frames of Trevor. With loop filter.

Appendix C. Presentation of Results on Tape

All the results presented on the demonstration tape were generated from the coded bit-streams. Since the images were re-generated from the bit-stream, cell-loss experiments can be easily performed by building some basic intelligence into the decoder. The single layer coder uses a loop filter, while the layered coder does not.

Sequences comparing layered coder versus single layer coder

The arrangement of video sequences for display is as follows:

Original (CIF)	Coder operating in layered (two layer) mode. Decoded Layer 1 image (CIF)
Coder operating in single layer mode Decoded CIF image.	Coder operating in layered (two layer) mode. Decoded Layer 0 image (QCIF)

Tables 1 and 2 provide detailed SNR and bit-rate figures for the coders illustrated on the demonstration tape. Note that the CIF PSNR is measured with respect to the original sequence and the QCIF PSNR is measured with respect to the QCIF sequence obtained by sub-sampling the original CIF sequence.

Single Layer Mode (with loop filter)	Table Tennis (1 MBit)	Tempete (1 MBit)	Flower Garden (1 MBit)	Flower Garden (4 MBit)	Trevor (384kBit)
CIF PSNR (Luminance) Y	31.35 dB	29.59 dB	24.01 dB	31.63 dB	36.27 dB
CIF PSNR (Chrominance) U	36.41 dB	29.71 dB	27.22 dB	32.42 dB	41.03 dB
CIF PSNR (Chrominance) V	35.72 dB	31.94 dB	29.98 dB	32.97 dB	42.92 dB
Quantizer step size	9	12	22	8	8
Peak CIF bit-rate	5065 kbits/sec	3226 kbits/sec	3785 kbits/sec	10370 kbits/sec	2893 kbits/sec
Average bit-rate for CIF only	996 kbits/sec	995 kbits/sec	1078 kbits/sec	4010 kbits/sec	307 kbits/sec

Table 1. Statistics of Presentation on Tape. Average over 125 frames for all sequences except for Trevor which is only 75 frames.

Two Layer Mode	Table Tennis (1 MBit)	Tempete (1 MBit)	Flower Garden (1 MBit)	Flower Garden (4 MBit)	Trevor (384kBit)
CIF PSNR (Luminance) Y	30.94 dB	28.95 dB	22.88 dB	31.14 dB	35.03 dB
CIF PSNR (Chrominance) U	36.01 dB	29.33 dB	26.38 dB	32.28 dB	40.02 dB
CIF PSNR (Chrominance) V	35.13 dB	31.61 dB	29.43 dB	32.75 dB	41.68 dB

CIF Quantizer step size	9	13	25	8	8
QCIF PSNR (Luminance) Y	33.92 dB	29.69 dB	25.94 dB	26.33 dB	32.41 dB
QCIF PSNR (Chrominance) U	37.83 dB	30.27 dB	28.39 dB	28.62 dB	37.84 dB
QCIF PSNR (Chrominance) V	36.77 dB	32.31 dB	30.84 dB	31.04 dB	39.27 dB
QCIF Quantizer step size	7	11	17	16	16
Peak QCIF bit-rate	1850 kbits/sec	1299 kbits/sec	1607 kbits/sec	1698 kbits/sec	1698 kbits/sec
Peak Layer 1 bit-rate	3271 kbits/sec	2072 kbits/sec	2064 kbits/sec	8973 kbits/sec	8973 kbits/sec
Average bit-rate for QCIF only	308 kbits/sec	355 kbits/sec	319 kbits/sec	344 kbits/sec	81 kbits/sec
Average bit-rate for Layer 1 data only	712 kbits/sec	770 kbits/sec	686 kbits/sec	3503 kbits/sec	252 kbits/sec
Average rate for CIF and QCIF data	1020 kbits/sec	1125 kbits/sec	1005 kbits/sec	3847 kbits/sec	333 kbits/sec

Table 2. Statistics of Presentation on Tape. Averaged over 125 frames for all sequences except for Trevor which is only 75 frames.

Sequences with cell loss

A preliminary cell-loss experiment was conducted using the bit-stream from the coded sequence Trevor. The bit-stream is packetized into 44 byte cells. After packetization, cell loss is introduced using the Gilbert Model, to simulate a network operating at 70 - 80 % utilization. No cell sequence numbering or error correction is used. In the event of a cell loss, the whole 44x8 bits are replaced with 1's. Error detection is achieved by detecting invalid VLC codes. The average cell loss rate used was 10^{-2}. Cell loss was applied to the complete bit-stream from the single layer coder, and to the layer 1 bit-stream in the layered coder. The arrangement of the decoded sequences is as follows:

Decoded Layer 1 (CIF) image with no cell loss	Coder operating in two layer mode with cell loss @ 10^{-2}. No Error Concealment.
Coder operating in single layer mode with cell loss @ 10^{-2}.	Coder operating in two layer mode with cell loss @ 10^{-2}. Error Concealment Strategies.

Note that "No error concealment" implies that the decoder assumes all information is lost (including the lower layer) when a cell is lost. This allows the areas in which cell loss occurs to be identified. The layered coder conceals errors by using the lower layer information alone to reconstruct the CIF resolution image.