INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11
MPEG2002/M8672
Klagenfurt, July 2002

Title: Spatio-Temporal Scalability in DCT-based Hybrid Video Coders

Source: Lukasz Blaszak, Marek Domanski, Adam Luczak, Slawomir Mackowiak
Poznan University of Technology, Poznan, Poland
Contact: M. Domanski (domanski@et.put.poznan.pl)
Group: MPEG-4
Subgroup: Video
Purpose: Proposal

1. Introduction

MPEG-4 provides Fine-Granularity-Scalability (FGS) for precise matching the bitstream to channel capacity. The lack of motion-compensated temporal prediction in the enhancement layer ensures that no accumulating error occurs when only a portion of the enhancement bitstream is received by the decoder. Moreover, the bitrate of the received bitstream part can be seamlessly controlled in a relatively easy-to-implement system. Unfortunately, the MPEG-4 FGS coders exhibit significant scalability overhead as compared to respective single-layer (non-scalable) coders.

Current research activities in scalability are related to two major groups of approaches [1]: wavelet-based techniques and improvements of the hybrid transform coders. The difficulty with the first approach is related to the necessity of embedding motion-compensated prediction into wavelet-based systems. Nevertheless, the 3-D-wavelet approach seems to be very promising as it provides inherently scalable systems. The second approach [2-8] mostly exploits multi-loop systems. Here, we are using the latter approach. Some similar approaches are those described in [3,8].

Prospective applications of scalable video coding include, e.g., wireless transmission. A practical requirement is that the base layer betrate is lower than the enhancement layer bitrate. In order to meet this requirement for two-layer systems, it has been proposed to combine the spatial scalability with other types of scalability thus reducing the base layer bitrate. The exemplary solutions are the following:

- combination of spatial and SNR scalability [e.g. 8],
- combination of spatial and temporal scalability called spatio-temporal scalability [6,9]. Here the base layer represents a video sequence with reduced both temporal and spatial resolutions. This contribtion exploits the latter approach.

The features of this proposal are:

- independent motion estimation and compensation in both loops,
- mixed spatio-temporal scalability,
- modified coding of B-frames.

In this contribution, a scalable video coder is proposed that consists of standard MPEG-4 building blocks. Moreover, the bitstream syntax is fully standard in the base layer and it needs only one modified macroblock-related field in the enhancement layers.

2. Coder structure

The scalable coder proposed consists of two (or more) motion-compensated coders (Fig. 1) that encode a video sequence and produce two bitstreams corresponding to two different levels of both spatial and temporal resolution (Fig. 1). Each of the coders has its own prediction loop with own motion estimation. Experiments have proved that the optimum motion vectors are different at different spatio-temporal resolutions (Fig. 2). The experimental data suggest

that often the bitrate needed for additional motion vectors is well compensated by the decease in the number of bits spent for the transform coefficients needed for prediction error encoding [5].

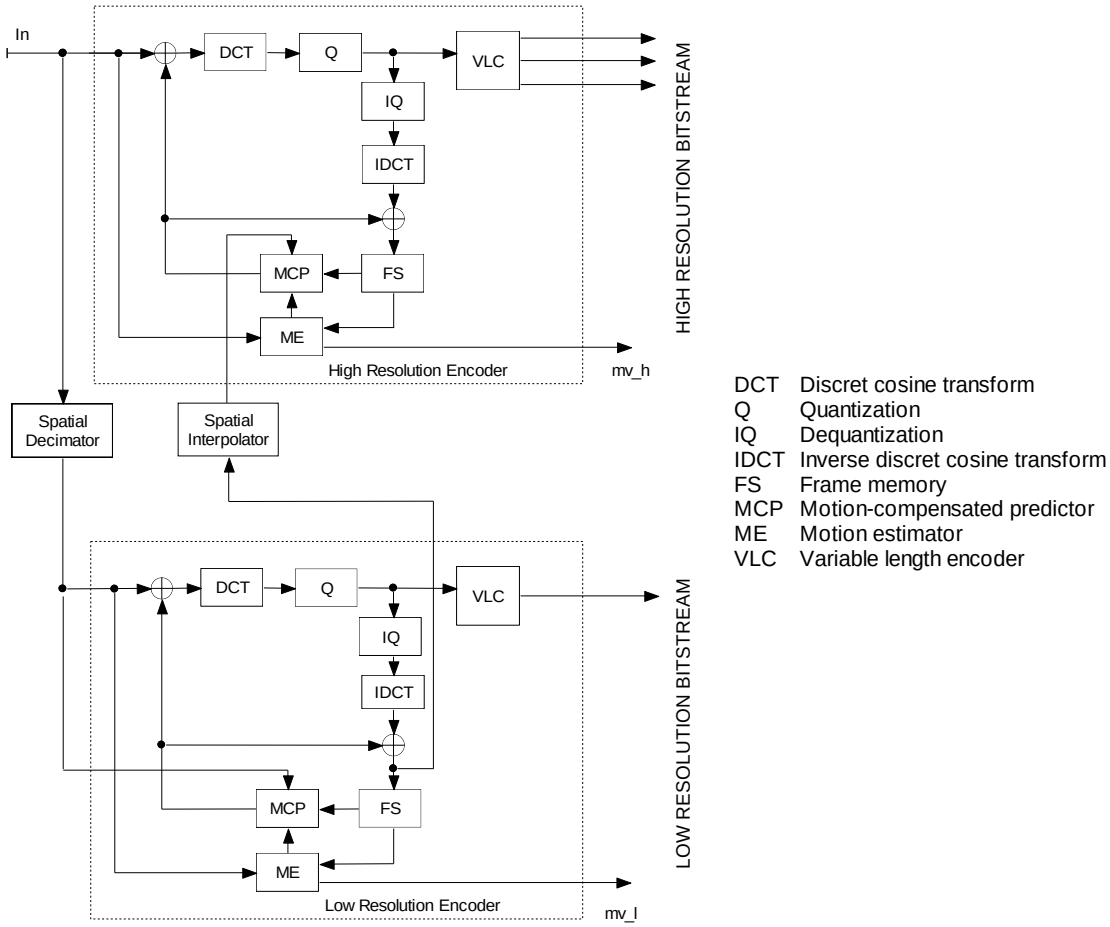

Fig. 1. General coder structure.

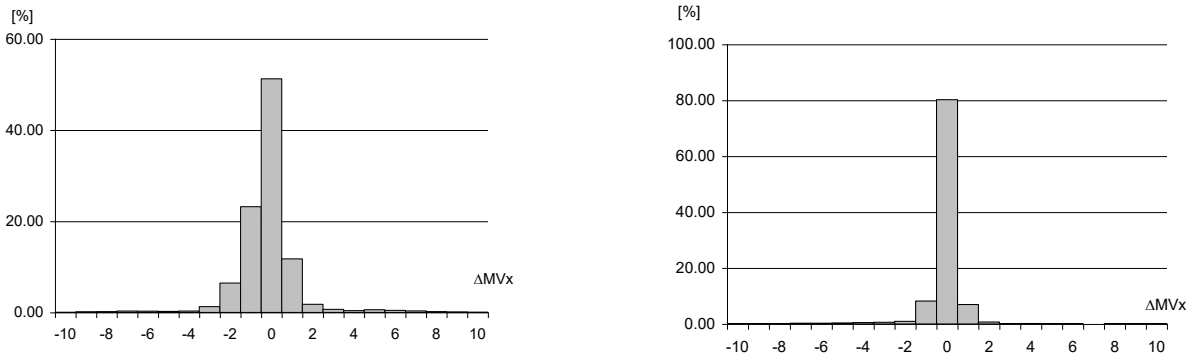

Fig. 5. Typical histograms of rounded differences between the base and enhancement layer motion vectors. The two plots correspond to two vector components.

Fine granularity may be obtained by data partitioning in the enhancement layer. Various schemes may be applied:

- successive transmission of bitplanes from the individual blocks,
- successive transmission of the consecutive nonzero transform coefficients, i.e., the (run,level) pairs.

In that way, the bitstream fed into a channel may be well matched with the throughput available. It means that the decoding process exploits only a part of one bitstream thus suffering from drift. Always, only one of the bitstreams is split, usually the medium- or high-resolution one. Therefore only one of the bitstreams received is affected by drift.

3. Video sequence structure

The phenomenon of drift is related to the reconstruction errors which are accumulating during the process of decoding of the consecutive frames. Therefore insertion of intra-coded frames bounds propagation of drift errors to groups of pictures (GOPs) (Fig. 3). In the enhancement layer, the additional I-frames have the bitsream syntax of P-frames with no motion vectors. Actually, all macroblocks in such a frame can be predicted from the interpolated low-resolution frames from the base layer.

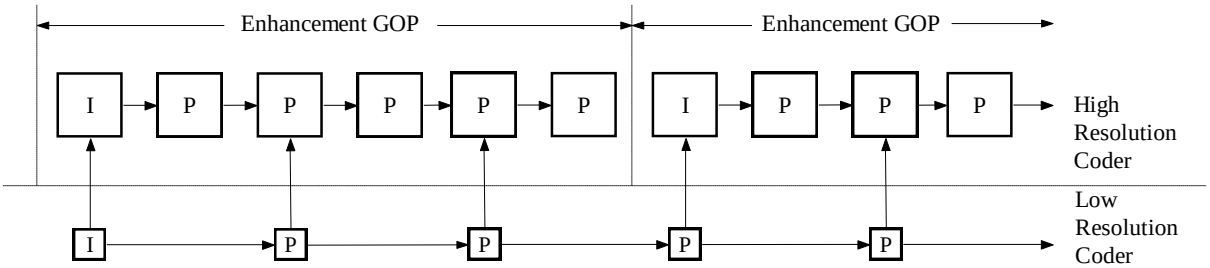

Fig. 3. Exemplary structure of a video sequence: There exist GOP structures in the enhancement layer where the I-frames are encoded with respect to the interpolated I- or P-frames from the base layer.

Moreover, higher percentage of B-frames also causes that drift accumulates slower. In the case of sequences with B-frames, quit efficient means of temporal decimation can be used, i.e., skipping some or all B-frames.

For higher percentages of B-frames in a sequence, there exist two types of B-frames: BE-frames processed by the high-resolution coder only and BR-frames that are processed by both coders like I- and P-frames. An exemplary but typical enhancement GOP structure is as follows:

I–BE–BR–BE–P–BE–BR–BE– P–BE–BR–BE– P–BE–BR–BE.

The full-resolution BR-frames may be used as reference frames for temporal prediction in the enhancement layer. Such a video sequence structure corresponds to a relatively low propagation of drifting errors.

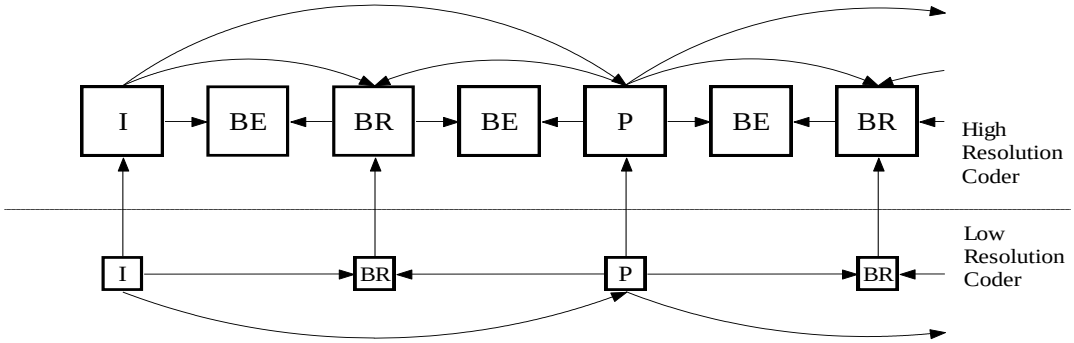

Fig. 4. Exemplary structure of a video sequence: Number of B-frames is 75% of the total number of frames. The BR-frames used as reference frames for the BE-frames in the enhancement layer.

4. B-frame encoding

Improved prediction is used for the BR-frames, which are the B-frames represented in both layers. Each macroblock in a full-resolution BR-frame can be predicted from the following reference frames (Fig.5):

- previous reference frame RP (I- or P-frame),
- next reference frame RN (I- or P-frame),
- current reference frame RC (BR-frame).

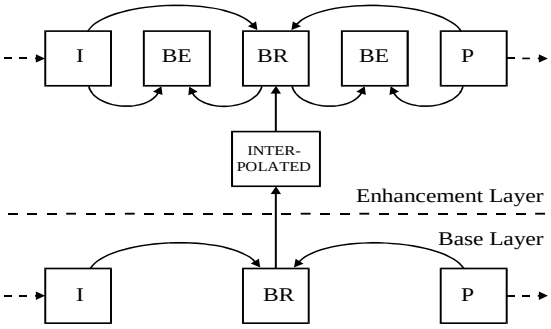

Fig. 5. Improved prediction of B-frames.

The data from the previous and next reference frames RP and RN are motion-compensated, and data from the current reference frame are upsampled in the two-dimensional space domain. The best suited reference frame or average of two or three reference frames is chosen according to the criterion of smallest prediction error. The improvement consists in another decision strategy. The best prediction/interpolation is chosen from all three possible reference frames: previous, future and interpolated (Fig. 5).

Experimental results prove that the current reference frame, i.e. a low-resolution BR-frame, is used in prediction of significant portion of macroblocks, sometimes even for more than 50% of all macroblocks in BR-frames (Fig. 6). Application of the above scheme of prediction leads to up to 20% lower bitrates for the enhancement-layer B-frames as compared to the scheme where the choice is made between spatial interpolation and the best temporal prediction.

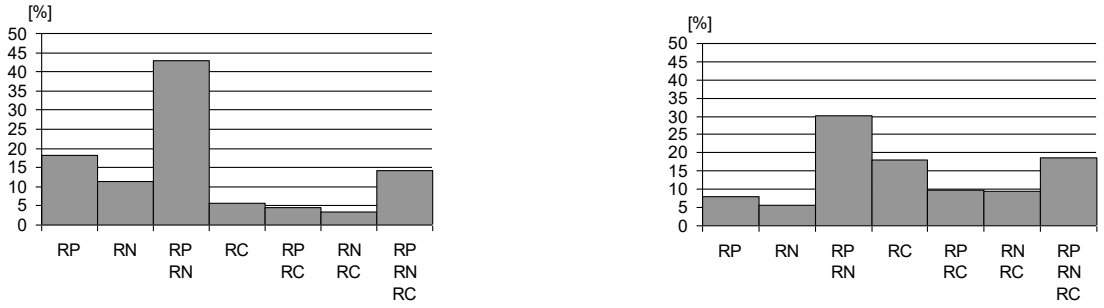

Fig. 6. Exemplary but typical percentages of macroblocks predicted using individual reference frames or their averages.

5. Bitstream structure

The base-layer bitsream is the standard single-layer MPEG bitsream.

In this contribution, in the enhancement high-resolution layer, we propose to encode the macroblock prediction type by use of variable-length codes. These codes could replace the COD flag in the macroblock headers.

For the case of P-frames, the estimated codes would be as in Table 1. The codes have been obtained from the statistics measured for three test CIF sequences at two bitrates.

Table 1. Exemplary codes for macroblock prediction type in P-macroblock headers.

Code	Reference	Data encoding (DCT coefficients and/or MVs)
1	temporal	yes
00	average	yes
010	temporal	no
0110	average	no
01110	interpolated	yes
01111	interpolated	no

The above field corresponds to the only bitstream syntax correction needed.

6. Experimental results

The structure has been tested using for some reference platforms and wide range of bitrates. .

The results obtained for the range of few megabits per second are quite promising as the bitrate overhead due to scalability is mostly below 10% with an MPEG-2 reference encoder. In these experiments, we used the sequences of structure shown in Fig. 4.

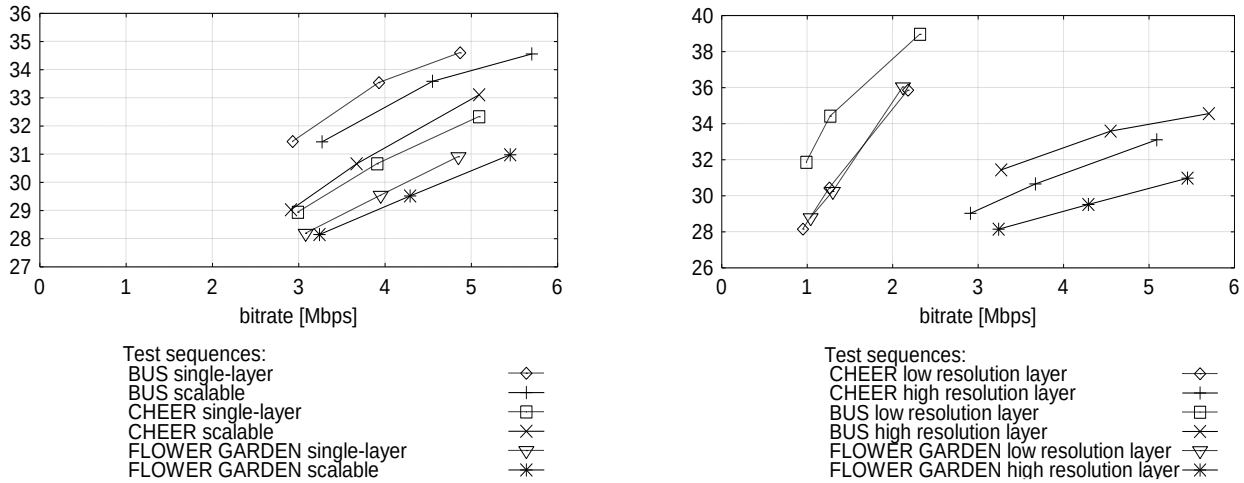

Fig. 7. Coding efficiency (luminance PSNR versus bitrate) for the two-loop structures (temporal subsampling factor = 2, image format = 4CIF, 50 fps). Scalable coders compared to the respective MPEG-2 single-layer coders.

For the H.263 reference codec, the bitrate overhead due to scalability has been measured relative to nonsacalble bitsream. This relative overhead depends strongly on the options switched on and on the quality of motion estimation as well. For example, the results for full-pel motion estimation exhibit sometimes even negative overhead for constant bitrate mode in both layers independently controlled (Fig. 8). Similar phenomena are also described in the references [10].

Improvement of the temporal prediction results in less intensive use of the reference frames interpolated from the base layer (Fig. 9). The bitrate overhead increases as shown in Table 2.

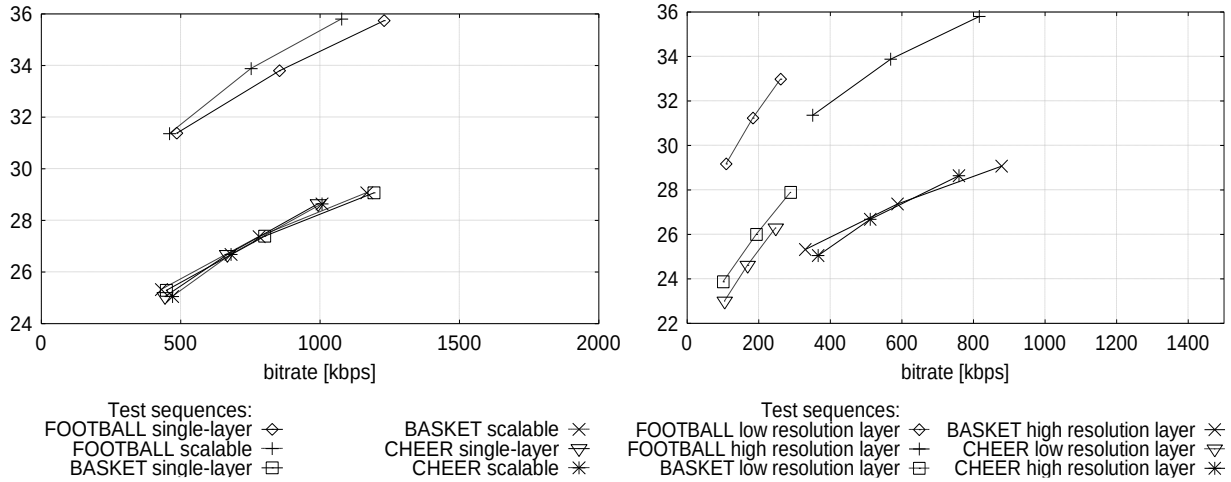

Fig. 8. The experimental results for the H.263 reference coder (full-pel resolution, motion vector search range ±16, independent bitrate control for both layers, and 4 MVs per macroblock enabled). No B-frames used.

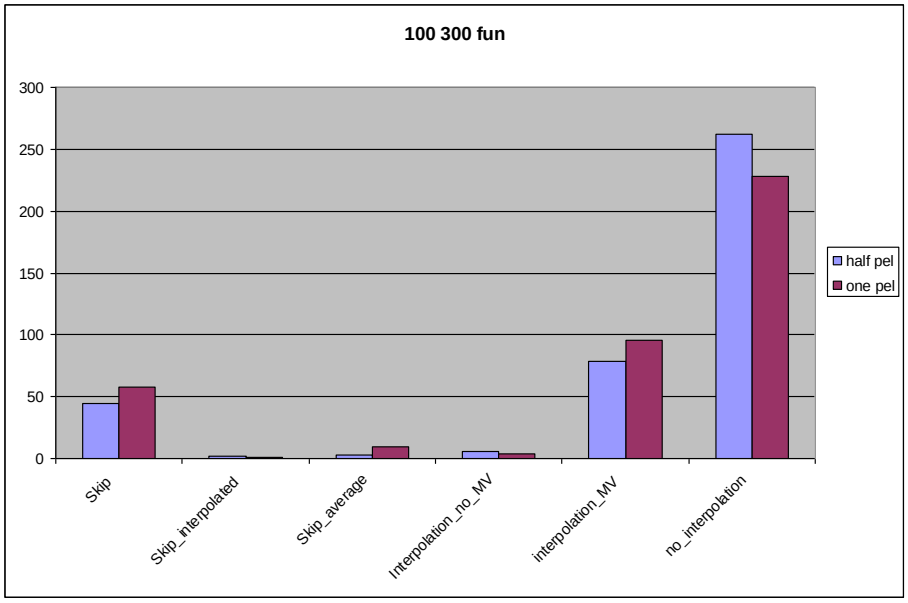

Fig. 9. Frequency of individual prediction modes for P-macroblocks measured for one- and half-pel motion estimation and compensation (Test sequence *Funfair* about 400 kbitps).

Table 2. Experimental results for half-pel motion estimation and compensation. No B-frames used.

H.263 - based coder for CIF (352 × 288) sequences		*Football*	*Basket*
Single-layer coder (H.263)	Bitstream [Kbps]	402.10	340.40
	Average luminance PSNR [dB]	31.61	25.63
Proposed scalable coder	Low resolution layer bitstream [Kbps]	109.91	100.31
	Low resolution layer average PSNR [dB] for luminance	30.10	24.71
	High resolution layer bitstream [Kbps]	356.98	329.30
	Average PSNR [dB] for luminance recovered from both layers	31.67	25.63
	Bitstream overhead [%]	16.11	26.21
Single-layer coder (H.263)	Bitstream [Kbps]	651.74	588.66
	Average luminance PSNR [dB]	34.76	28.13
Proposed scalable coder	Low resolution layer bitstream [Kbps]	186.89	194.52
	Low resolution layer average PSNR [dB] for luminance	32.78	27.28
	High resolution layer bitstream [Kbps]	566.51	584.50
	Average PSNR [dB] for luminance recovered from both layers	34.69	28.19
	Bitstream overhead [%]	15.60	32.33
Single-layer coder (H.263)	Bitstream [Kbps]	925.46	913.13
	Average luminance PSNR [dB]	36.70	30.15
Proposed scalable coder	Low resolution layer bitstream [Kbps]	262.95	289.56
	Low resolution layer average PSNR [dB] for luminance	34.61	29.23
	High resolution layer bitstream [Kbps]	815.45	878.41
	Average PSNR [dB] for luminance recovered from both layers	36.74	30.12
	Bitstream overhead [%]	16.53	27.91

The TML-9.4 coder (H.26L) uses more efficient temporal prediction that of H.263. It results in higher scalability overhead mostly not much lower than for simulcast (Fig. 10). The experiments have been made with default options in TML-9.4 for the CIF test sequences.

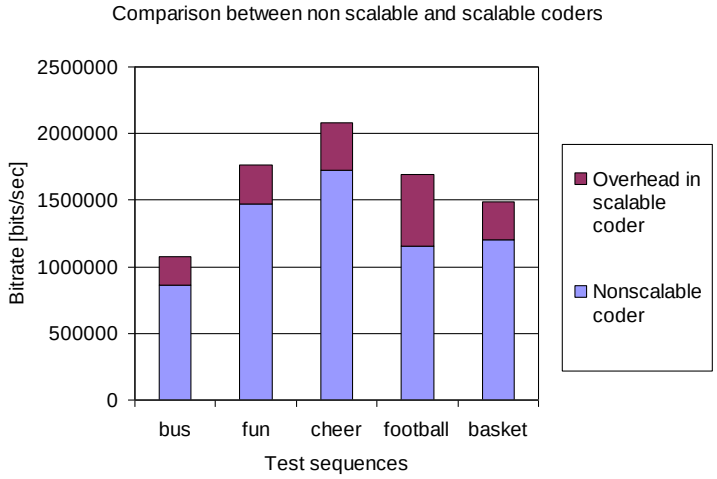

Fig. 10. A comparison between non scalable and scalable TML 9.4 – based coders.
Results obtained for progressive CIF sequences encoded with no B-frames.

For all the cases above considered, the base-layer was about 30- 40% of the total bitrate for all three decimation factors (temporal, horizontal and vertical) set to 2. Exemplary results for fine granularity scalability are shown in Figs. 11 and 12.

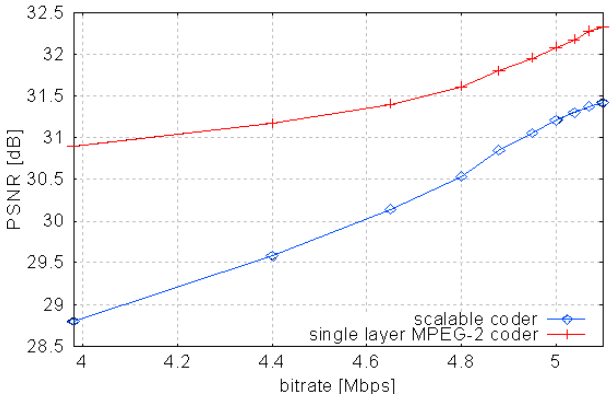

Fig. 11. Bitrate control using FGS proposed with the GOP length of 12.

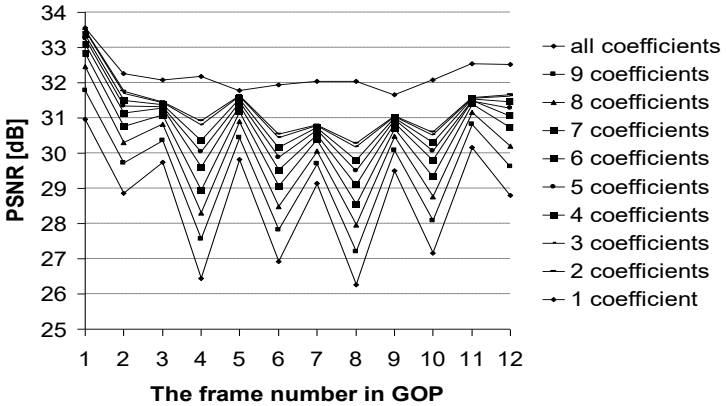

Fig. 12. Decreasing signal-to-noise ratio according to drift for various numbers of DCT coefficients per block transmitted in the enhancement layer to the decoder. The two-loop coder and the sequence structure from Fig. 4. Test

sequence *Funfair* with an average bitrate 5Mbps.

7. Conclusions

Described is a generic multi-loop coder structure for motion-compensated fine-granularity scalability. The basic features of the coder are:
- mixed spatio-temporal scalability,
- independent motion estimation for each motion-compensation loop, i.e. for each spatio-temporal resolution layer,
- BR/BE-frame structure,
- improved prediction of BR-frames.

For most experiments, the lost of quality due to fine granularity is not dramatic. In many applications, the reduced bitrate corresponds to some perturbation and therefore a certain loss of coding performance can be acceptable when the whole coder performs well. Drift can be reduced by dividing a particular layer bitstream into GOPs.

Unfortunately, for the TML-9.4, the coding performance is only slightly better than that of simulcast. On the other hand, the encoding time is not much higher than for simulcast. Some experiments show that coding efficiency can be improved using some smoothing of motion field and joint encoding of motion vectors from both resolution levels (both layers).

References

[1] J.-R.Ohm, M. Beermann, „Status of scalable technology in video coding“, Doc. ISO/IEC JTC1/SC29/WG11 MPEG01/M7483, Sydney, July 2001

[2] J.-R.Ohm, M. van der Schaar, *Scalable Video Coding*, Tutorial material, *Int. Conf. Image Processing ICIP 2001*, IEEE, Thessaloniki, October 2001.

[3] Y. He, R. Yan, F. Wu, S. Li, “H.26L-based fine granularity scalable video coding,” Doc. Doc. ISO/IEC JTC1/SC29/WG11 MPEG01/m7788, December 2001.

[4] K.Rose, S. Regunathan, “Toward optimality in scalable predictive coding,” *IEEE Trans. Image Proc*., vol. 10, pp.965-976, July 2001.

[5] M. Domański, S. Maćkowiak, „Modified MPEG-2 video coders with efficient multi-layer scalability,” *Int. Conf. Image Processing ICIP 2001*, IEEE, vol. II, pp. 1033-36, Thessaloniki, October 2001.

[6] M. Domański, A. Łuczak, S. Maćkowiak, „Spatio-temporal scalability for MPEG video coding”, *IEEE Trans. Circ. and Syst. Video Technology*, vol. 10, pp. 1088-1093, Oct. 2000.

[7] A. Reibman, L. Bottou, A. Basso, “DCT-based scalable video coding with drift,” *Int. Conf. Image Processing ICIP 2001*, IEEE, vol. II, pp. 989-992, Thessaloniki, October 2001.

[8] U. Benzler, “Spatial scalable video coding using a combined subband-DCT approach”, *IEEE Trans. Circuits and Systems for Video Technology*, vol. 10, pp. 1080-1087, October 2000.

[9] M. Domański, A. Łuczak, S. Maćkowiak, R. Świerczyński, ”Hybrid coding of video with spatio-temporal scalability using subband decomposition,” in *Signal Processing IX: Theories and Applications, Proc. EUSIPCO-98*, pp. 53-56, Rhodes, September 1998.

[10] G. Cote, B. Erol, M. Gallant, F. Kossentini, “H.263+: video coding at low bit rates”, *IEEE Trans. Circ. and Syst. Video Technology*, vol. 8, pp. 849-865, November 1998.