

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

**ISO/IEC JTC1/SC29/WG11
MPEG2003/M9895
Trondheim, July 2003**

Title: Modified AVC Codecs with Spatial and Temporal Scalability

Source: Łukasz Błaszak, Marek Domański
Poznań University of Technology, Poznań, Poland

Contact: M. Domański (domanski@et.put.poznan.pl)

Group: MPEG-4

Subgroup: Video

Purpose: Proposal

1. Introduction

The AVC Version 1 video codec does not support scalability that is currently considered as an important functionality for many applications, e.g., wireless systems with bandwidth variations and fadings, video broadcasting in heterogeneous communication networks, unequal error protection etc. [1,2]. There are many proposals for scalable coding techniques that can be used in the AVC coders [3]. Some approaches exploit wavelet analysis and synthesis while others try to embed scalability into the hybrid video codec structures and bitstream syntax [4,5].

In particular, the document M9469 [5] described a proposal for embedding spatial and temporal scalability into the AVC codec. For that proposal, an important assumption was made that the scalable codec had the same coding technology as already defined AVC codec and only minor modifications were imposed onto the codec structure. It means that a scalable AVC codec exploits mostly the same functional blocks as standard AVC codec. Moreover the bitstream syntax of all layer is the standard bitstream AVC syntax with minor modification of semantics. Also that the scalable AVC codec exhibits similar complexity as a simulcast cluster of AVC codecs.

The features of that proposal were:

- mixed spatio-temporal scalability,
- two motion compensation loops for two bitstreams with different spatio-temporal resolutions,
- independent motion estimation and compensation in both loops,
- additional reference frames in the enhancement layer sub-codec (the frames obtained using interpolation from the low-resolution base-layer reconstructed frames),
- some B-frames used as reference frames in the enhancement layer.

In this contribution, some modifications and extensions of the codec structure form [5] are described.

2. Multi-loop scalable coder structure

The scalable coder proposed consists of some motion-compensated coders that encode a video sequence and produce bitstreams corresponding to different levels of spatio-temporal resolutions. For example, a three-layer video representation may be produced by a three-loop video coder (Fig. 1). Such a coder is compliant with Test 1 of the *Call for Evidence on Scalable Video Coding Advances* [6].

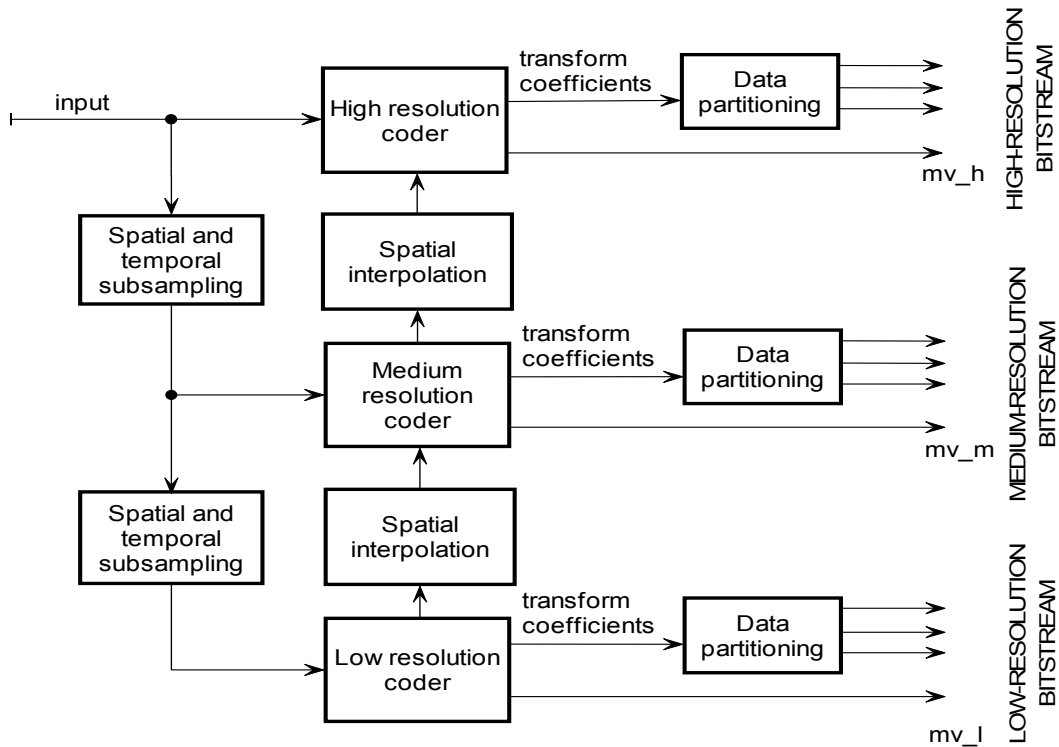


Fig. 1. A three-loop scalable coder.

Each of the coders has its own prediction loop with own motion estimation. Each sub-coder produces a bitstream that consists of two major parts:

- encoded transform coefficients,
- motion vectors .

The characteristic feature of this structure is independent motion estimation in both sub-coders resulting in optimum motion vectors estimated for all resolution levels. These motion vectors allow exact motion-compensated prediction in all layers.

In any enhancement-layer high-resolution sub-coder additional reference frames can be used for both backward and forward prediction, i.e. interpolated frame from the current low-resolution base-layer frame and linear combinations (averages) of the current interpolated frame and temporal reference. For the latter, independent motion estimation can be performed aiming at estimation of the optimum motion vectors that yield the minimum prediction error for the reference being an average of spatial and temporal references.

Fine granularity may be obtained by use of splitting the data produced on any resolution level. In that way, the bitstream fed into a decoder 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 high-resolution one. Therefore only one of the bitstreams received is affected by drift that is related to the reconstruction errors which are accumulating during the process of decoding of the consecutive frames. In this bitstream, inserted are I-frames encoded with respect to interpolated lower-resolution frames. In fact, such frames are encoded similarly as P-frames but with simultaneous interpolated reference frames. The bitstream syntax is that of P-frames. Insertion of such frames does not affect performance so much but bounds propagation of drift errors to groups of pictures (GOPs). Moreover, higher percentage of B-frames also decreases the influence of drift. The GOP length should be chosen in such a way that drift is acceptable in the worst case of lowest bitrate in the enhancement layer.

The choice of the spatial decimator and interpolator has substantial impact on the overall coding efficiency as described in [5]. In particular adaptive interpolation proved to give promising results.

3. Video sequence structure

In this section, we report the considerations from [5]. For the sake of simplicity, two-layer arrangements are considered. The generalization to three-layer arrangement is straightforward.

Time decimation is performed by frame skipping. The B-frames would be usually skipped if they exist (Fig. 2 a,b). Therefore there may exist two types of B-frames:

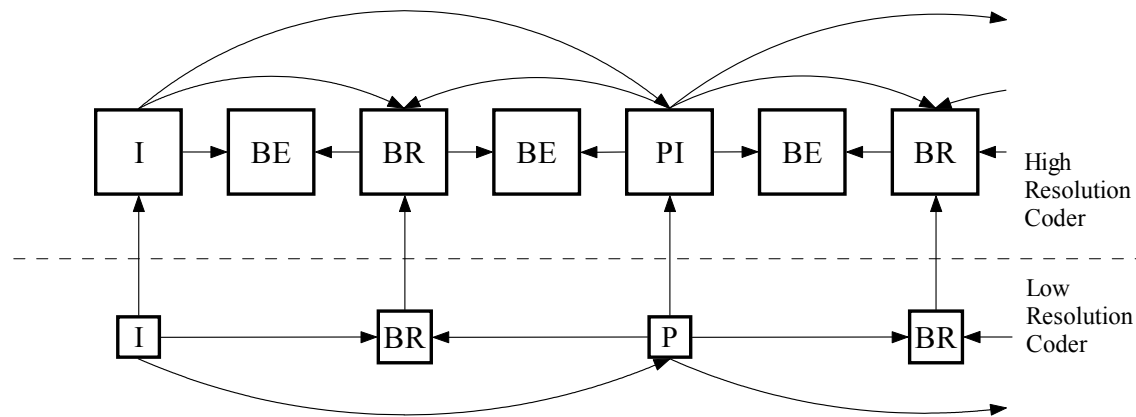
- BE-frames that exist in the enhancement layer only and
- BR-frame that exist both in the base and in the enhancement layer.

The latter may be predicted using interpolation from the decoded low-resolution base-layer frames (Fig. 2a).

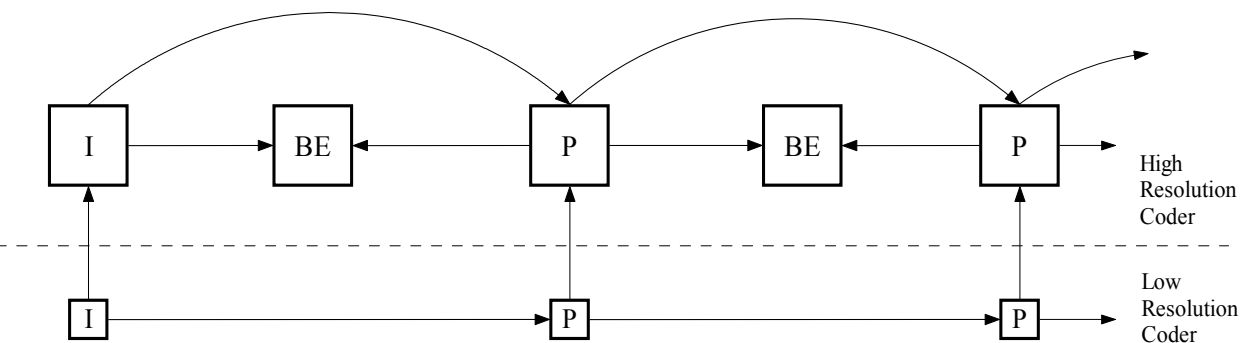
Typical temporal decimation factor values are 2 and 3.

In this proposal, improved B-frame encoding [7-10] was used in the enhancement layer, i.e. the BR-frame was used as a temporal reference for the neighboring BE-frames (Fig. 2a).

a)



b)



c)

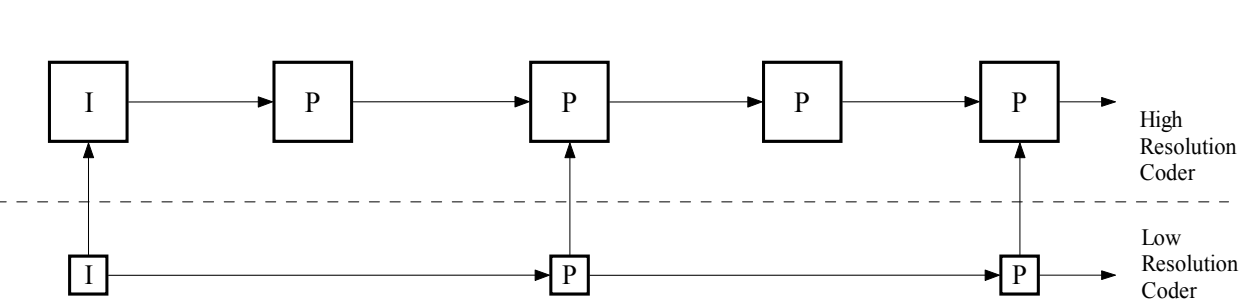


Fig. 2. Exemplary structures of low-resolution and high-resolution video sequences with temporal subsampling by factor 2.

4. Prediction modes in the high-resolution enhancement-layer sub-coder

Sophisticated intra- and interframe predictions are related to major performance improvements in the AVC coders. The enhancement-layer sub-coder employs additional prediction modes that exploit the current interpolated base-layer frame as the reference. Other modes exploit averages of temporal prediction and spatial interpolation as references. These modes are carefully embedded into the mode hierarchy of the AVC coder thus obtaining the binary codes that correspond to the mode probabilities. The respective mode hierarchy is shown in Table 1 [5].

The choice of the lowest-cost prediction mode plays the key role. The encoding scheme would reduce to simulcast when no interpolated reference macroblocks are used in the enhancement layer. In the other extreme situation, in the enhancement layer, no temporal prediction is used, and only interpolated base-layer frames are used for prediction of the enhancement macroblocks (like in MPEG-4 FGS). The latter situation is very unlikely because of the high efficiency of the AVC temporal prediction. Nevertheless the extreme situations are related to unsatisfactory coding performance. The spatial interpolation must be very efficient in order to avoid them. Good fidelity of the decimation-interpolation scheme results in reasonable probability that the reference sample block interpolated from the base layer leads to smaller prediction error as compared to the temporal prediction within the enhancement layer.

Table 1. Prediction mode hierarchy (repeated from [5]).

Frame type	Prediction modes
Intra (I)	<ol style="list-style-type: none"> 1. Spatial interpolation from base layer (16×16 block size). 2. All standard intra prediction modes.
Inter (P)	<ol style="list-style-type: none"> 1. Prediction (forward) from the nearest reference frame. 2. Spatial interpolation from base layer (16×16 - 4×4 block size). 3. Average of two above (1, 2). 4. Temporal prediction modes from other reference frames in the order defined in AVC specification. 5. All standard intra modes.
Inter (B)	<ol style="list-style-type: none"> 1. Prediction (forward, backward and bidirectional) from the nearest reference frame. 2. Spatial interpolation from base layer (16×16 - 4×4 block size). 3. Average of two above (1, 2). 4. Temporal prediction modes from other reference frames in the order defined in AVC specification. 5. All standard intra modes.

The base-layer bitstream is the standard single-layer AVC bitstream. In the contribution [5], the interpolation modes are considered as prediction from an additional reference frame, i.e. the interpolated frame from the lower spatial resolution. Therefore the standard bitstream syntax may be used. The modifications of the semantics mean that the codes of the reference frames are modified according to the hierarchy from Table 1.

5. Modified averaging mode

In the former proposal [5], the averaging mode (see Table 1) corresponds to prediction of a macroblock as an average of:

- the block interpolated from the respective samples from the base layer,
- the block from the previous or future temporal reference that is pointed out by the appropriate motion vector.

This mode will be called “averaging mode 1”(Fig. 2).

Here, we propose to use an “averaging mode 2”. In the enhancement or middle layer, a predicted macroblock

as a sum (and not an average) of:

- the block interpolated from the respective samples from the base layer,
- the block from the previous or future high-pass filtered reference that is pointed out by the appropriate motion vector.

For both modes, the motion vectors estimated for temporal prediction. Nevertheless, new motion vectors may be calculated for this “averaged reference”. This approach improves slightly the performance, and is used both in [5] and here.

AVERAGING MODE 1 :

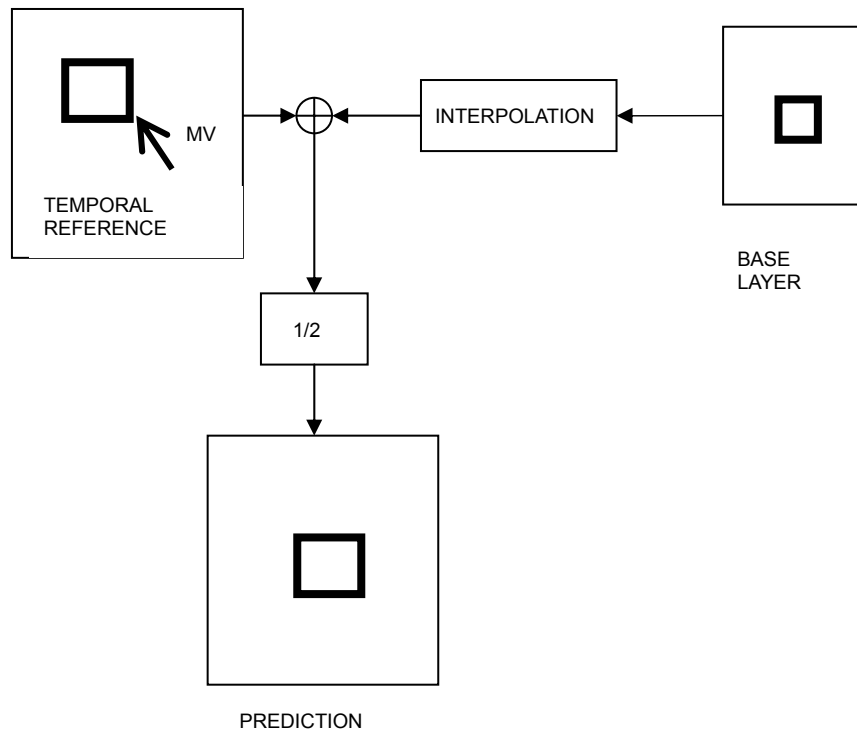


Fig. 3. Averaging mode 1.

AVERAGING MODE 2 :

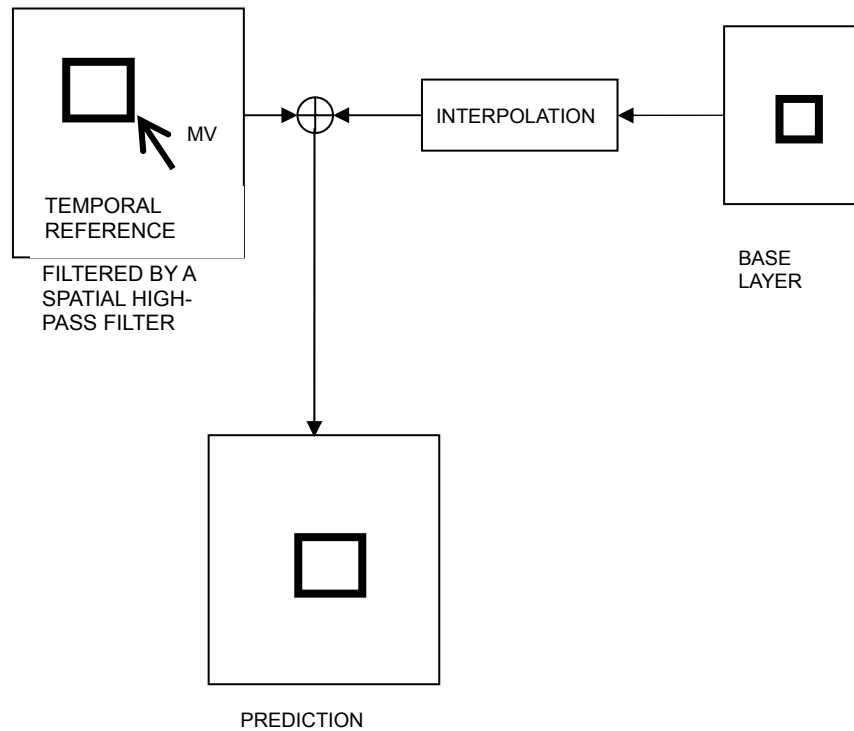


Fig. 4. Averaging mode 2.

The idea of this mode of prediction is close to motion-compensated subband synthesis.

6. Experimental results

The scalable test model has been implemented on the top of standard JVT software version 2.1. Both coder and decoder have been implemented as in [5]. The major difference with respect to [5] is that here the averaging mode 2 was used instead of mode 1 tested in [5].

In order to test the coding performance of the scalable AVC codec with the averaging mode 2, a series of experiments have been performed with (352×288) -pixel sequences. Horizontal, vertical and temporal subsampling factors have been set to 2 and the video sequence structure was that from Fig. 2a.

In the experiments, the following modes have been switched on:

- CABAC coder,
- $\frac{1}{4}$ -pel motion estimation in both layers,
- all prediction modes.

The experiments have been performed for three sets of the quantization parameter values. These values were defined independently for I-frames (QP_I), P-frames (QP_P) and B-frames (QP_B). In the tests, equal values of QP_I , QP_P and QP_B were applied in the base and the enhancement layer, respectively.

In order to compare the scalable codec with the non-scalable reference AVC codec as well as with the simulcast pair of non-scalable AVC codecs, the experiments have been performed with constant values of QP_I , QP_P and QP_B that imply almost constant quality measured in terms of the PSNR factor for the luminance component in a given sequence. Of course, the quality measured for different sequences is different, but for a given video sequence and a given set of QP_I , QP_P and QP_B , the results for scalable, non-scalable and simulcast coding differ mostly less than 0.3 dB and often even less than 0.1 dB. For such conditions, bitrates have been estimated for the scalable coder (*whole scalable coder*), non-scalable coder and simulcast coding (Table 2 and Fig. 5).

For such test conditions, the approximate bitrate overhead due to scalability was between -1% and 12% of

the bitrate for the nonscalable (single-layer) codec. For almost all cases, scalable coder performed better than simulcast coding. Usually scalable coding performance was substantially higher than that of simulcast.

Within a scalable coder, the base layer bitrate was about 15% to 22% of the total bitrate produced by a scalable coder for both layers.

The results show that coder performance is similar for both averaging prediction modes 1 and 2. The prediction mode 2 needs slightly more computations.

Table 2. Coding efficiency comparison for scalable, nonscalable and simulcast coding of (352 × 288) sequences. Horizontal, vertical and temporal subsampling factors are set to 2. Sequence structure from Fig. 2a is applicable.

QI = 10, QP = 11, QB = 12.

	Basket(25Hz)		Bus(30Hz)		Cheer(30Hz)		Football(30Hz)		Fun(25Hz)	
	PSNR [dB]	kbit/s	PSNR [dB]	kbit/s	PSNR [dB]	kbit/s	PSNR [dB]	kbit/s	PSNR [dB]	kbit/s
Base layer	37.21	494.78	38.65	277.91	37.79	726.90	39.75	353.11	37.70	465.20
Enhancement layer	38.09	2201.08	39.03	1517.22	38.66	2969.77	40.74	1392.10	38.58	2105.08
Non scalable	38.06	2506.54	39.10	1622.20	38.64	3512.30	40.74	1736.30	38.55	2312.69
Simulcast	38.06	3001.32	39.10	1900.11	38.64	4239.20	40.74	2089.41	38.55	2777.89
Scalable	38.09	2695.86	39.03	1795.13	38.66	3696.67	40.74	1745.21	38.58	2570.28
Overhead [%]										
Scalable	7.55		10.66		5.25		0.51		11.14	
Simulcast	19.74		17.13		20.70		20.33		20.12	

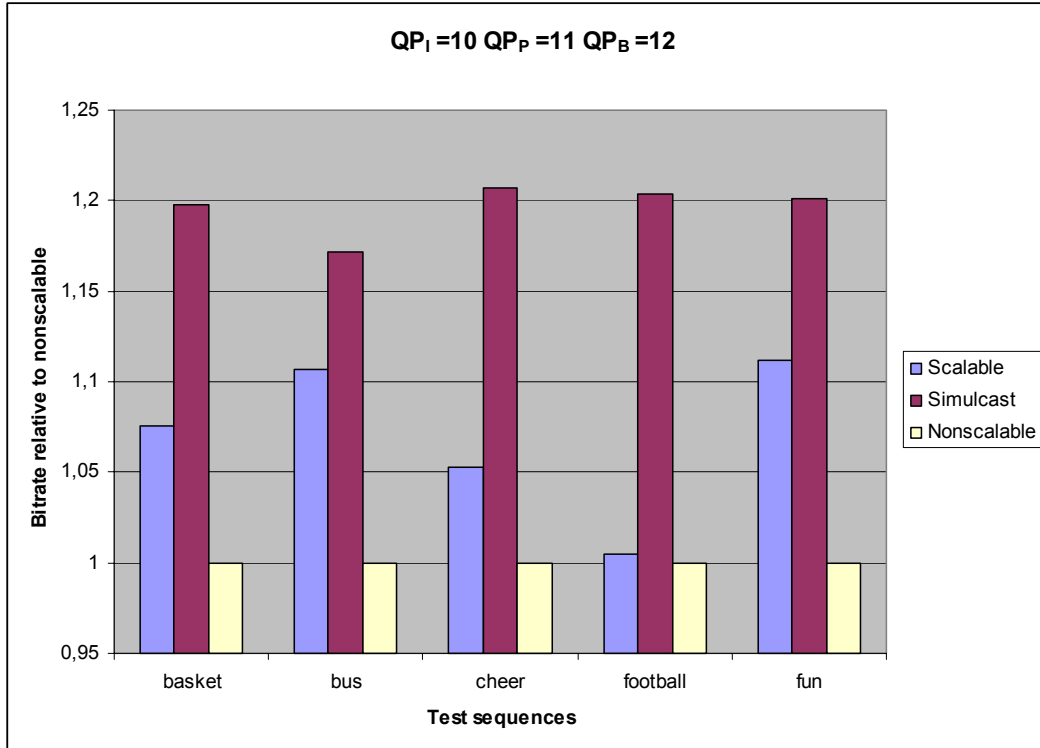
QI = 15, QP = 16, QB = 17.

	Basket(25Hz)		Bus(30Hz)		Cheer(30Hz)		Football(30Hz)		Fun(25Hz)	
	PSNR [dB]	kbit/s	PSNR [dB]	kbit/s	PSNR [dB]	kbit/s	PSNR [dB]	kbit/s	PSNR [dB]	kbit/s
Base layer	33.03	281.68	34.68	150.94	33.59	438.30	36.21	199.10	33.66	270.78
Enhancement layer	34.11	1179.82	34.84	805.71	34.61	1665.00	37.30	763.19	34.70	1186.88
Non scalable	34.13	1364.15	35.21	857.63	34.66	2017.29	37.44	968.01	34.74	1310.05
Simulcast	34.13	1645.83	35.21	1008.34	34.66	2455.59	37.44	1167.11	34.74	1580.83
Scalable	34.11	1461.50	34.84	956.65	34.61	2103.30	37.30	962.29	34.70	1457.66
Overhead [%]										
Scalable	7.14		11.55		4.26		-0.59		11.27	
Simulcast	20.65		17.60		21.73		20.57		20.70	

QI = 20, QP = 21, QB = 22.

	Basket(25Hz)		Bus(30Hz)		Cheer(30Hz)		Football(30Hz)		Fun(25Hz)	
	PSNR [dB]	kbit/s	PSNR [dB]	kbit/s	PSNR [dB]	kbit/s	PSNR [dB]	kbit/s	PSNR [dB]	kbit/s
Base layer	29.14	154.33	31.05	77.20	29.76	248.35	33.14	108.05	29.89	148.37
Enhancement layer	30.33	643.81	31.34	429.81	30.91	935.63	34.06	424.93	30.99	659.23
Non scalable	30.40	746.52	31.68	454.55	31.03	1150.07	34.38	543.62	31.09	727.56
Simulcast	30.40	900.85	31.68	531.75	31.03	1398.42	34.38	651.67	31.09	875.93
Scalable	30.33	798.14	31.34	507.01	30.91	1183.98	34.06	532.98	30.99	807.60
Overhead [%]										
Scalable	6.91		11.54		2.95		-1.96		11.00	
Simulcast	20.67		16.98		21.59		19.88		20.39	

a)



b)

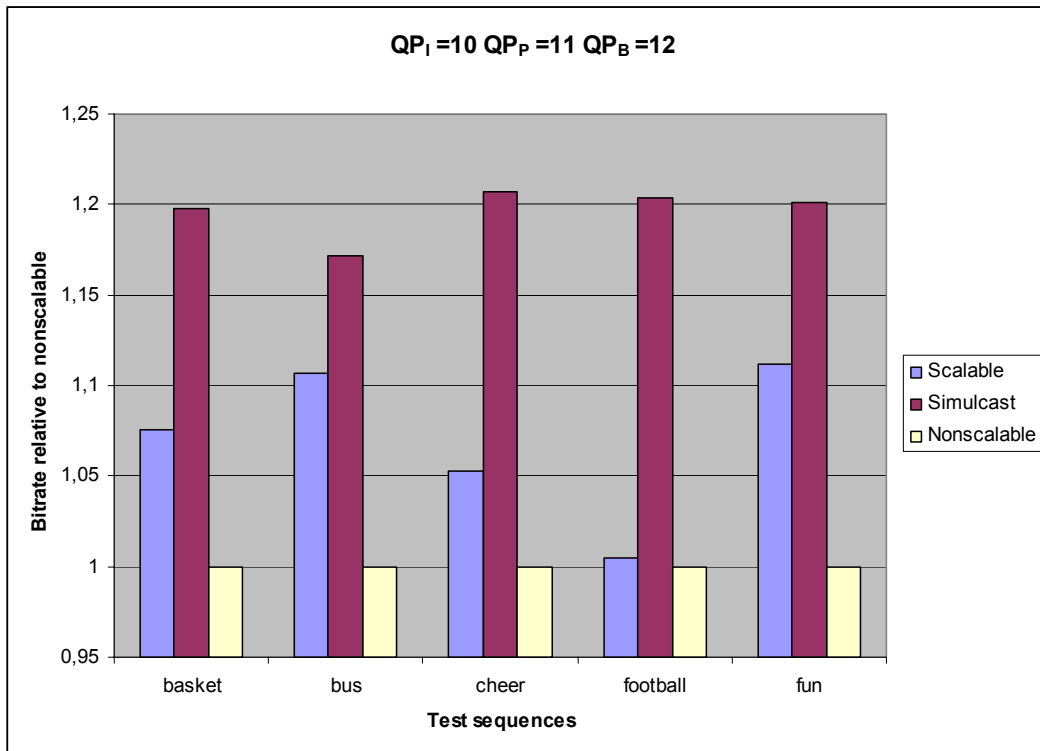


Fig. 5. Approximate bitrate comparison for scalable, nonscalable (single-layer) and simulcast coding.

c)

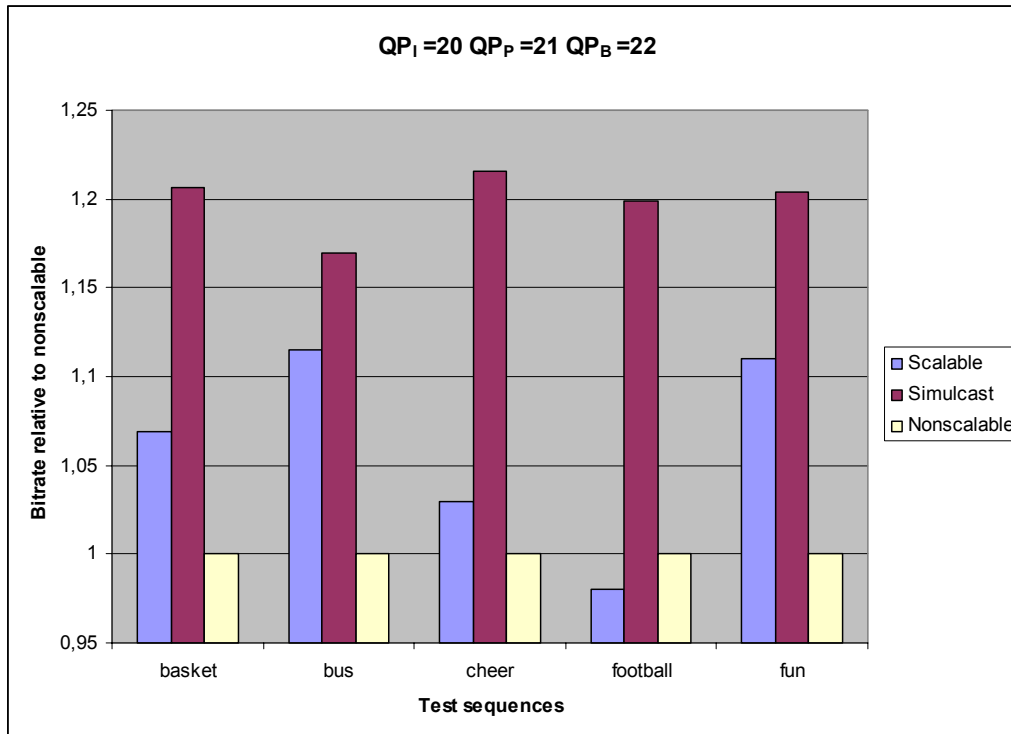


Fig. 5. (cont'd) . Approximate bitrate comparison for scalable, nonscalable (single-layer) and simulcast coding.

7. Response to Call for Evidence - Test 1a

In order to respond to [6] a new series of experiments have been performed that correspond to Test 1a . The bitstreams and the respective decoded sequences are available for subjective quality assessment.

7.1. Test sequences and bitrates

The test sequences were; *Harbor*, *Crew*, *Night*, *Sailormen* available at <ftp.tnt.uni-hannover.de/testsequences>. The sequences are in the resolution 720×480 pixels and 30 frame/s. The coding experiment consists of evaluation of the three-layered video coding using the AVC scalable video codec as proposed in [5] with the three-loop extension described in Sec. 2 of this document. Therefore, the usage of the averaging mode 1 was assumed.

7.2. Coder settings

Options :

- ME accuracy = 1/4 (for all layers)
- ME search range = 32 for SD , 32 for CIF, 16 for QCIF,
- CABAC,
- RD optimization on,
- loop filter - yes,
- there are no access frames,
- the full SD resolution sequence: IBBBPBBBBPBBBBP... ,
- the CIF sequence (middle layer): IBPBPBPBP ... ,
- the QCIF sequence (base layer): IBPBPBPBP ... ,
- number of reference pictures = 4 (by full temporal resolution).

Please note that for the SD resolution the motion vector search range was halved as compared to the AVC anchor conditions defined in Table 5 of [6].

Expected encoding and decoding delay: is the same as AVC with respect to the sequence structure of the lowest layer.

Expected computational complexity - versus standard single-layer AVC coder with the same options:

- Coder: approx. 1.3 times single-layer coder (for two layer representation with spatial scalability),
- approx. 1.45 times single-layer coder (for three-layer representation with spatial scalability),
- approx. 1.15 times single-layer coder (for two-layer representation with spatio-temporal scalability).

FGS was not tested.

The base-layer bitstream is fully compliant with AVC. Enhancement-layer bitstreams have slightly modified semantics with standard syntax of AVC.

7.3. Results

Table 3. The results for the test 1a (the first column of the table 1 [6]).

Sequence	Bitrate [kbps]	Y [PSNR]	U [PSNR]	V [PSNR]
Night_720x480_30Hz	1440.03	32.70	37.20	39.50
Night_360x240_15Hz	403.69	31.07	35.28	37.71
Night_180x120_7.5Hz	66.11	32.07	34.99	37.96
Crew_720x480_30Hz	1559.37	36.00	40.28	40.17
Crew_360x240_15Hz	406.73	34.88	38.41	37.40
Crew_180x120_7.5Hz	68.49	34.92	37.40	36.33
Harbour_720x480_30Hz	1508.53	31.16	40.38	42.48
Harbour_360x240_15Hz	298.51	27.47	38.61	41.26
Harbour_180x120_7.5Hz	66.95	31.32	39.38	41.61
Sailormen_720x480_30Hz	1631.30	33.88	37.71	38.70
Sailormen_360x240_15Hz	391.28	32.20	35.96	36.87
Sailormen_180x120_7.5Hz	65.21	35.09	37.37	38.05

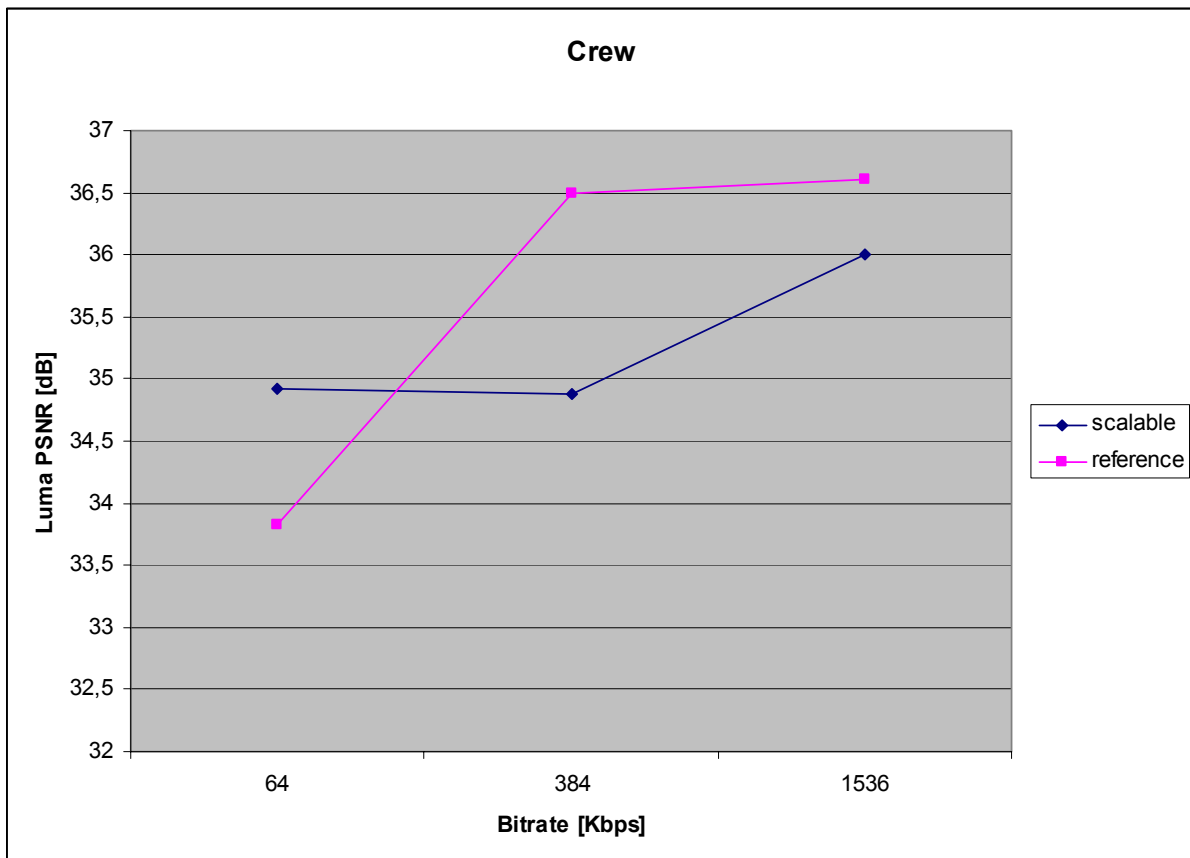


Fig. 6. The results for the test 1a (the first column of the table 1 [6]).

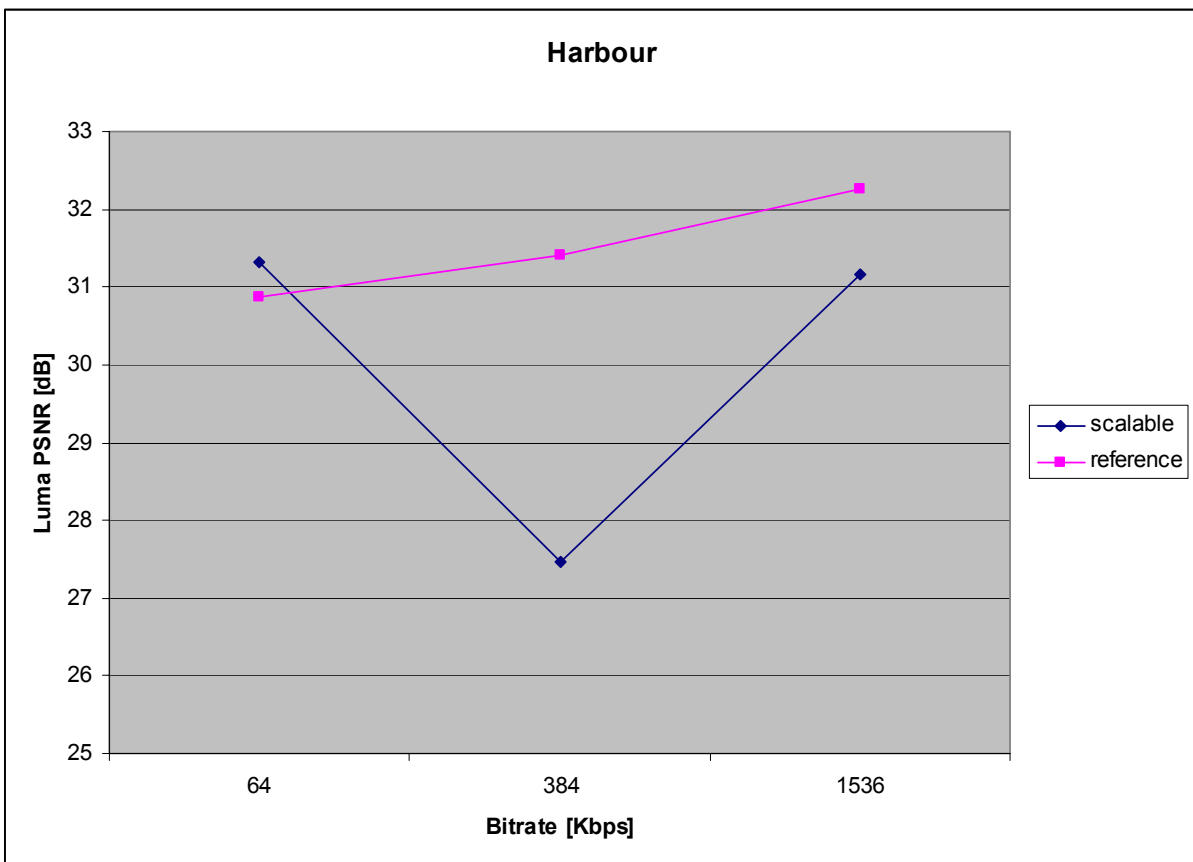
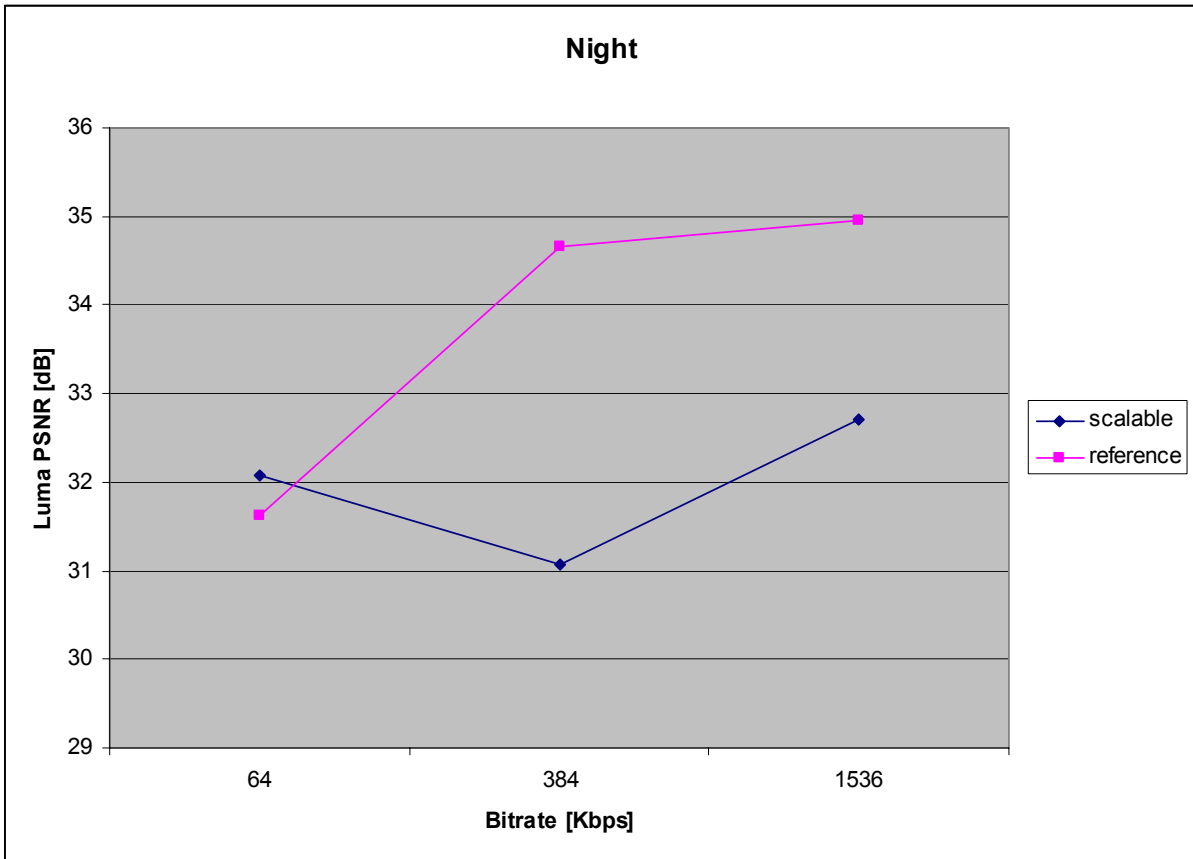


Fig. 6 (cont'd). The results for the test 1a (the first column of the table 1 [6]).

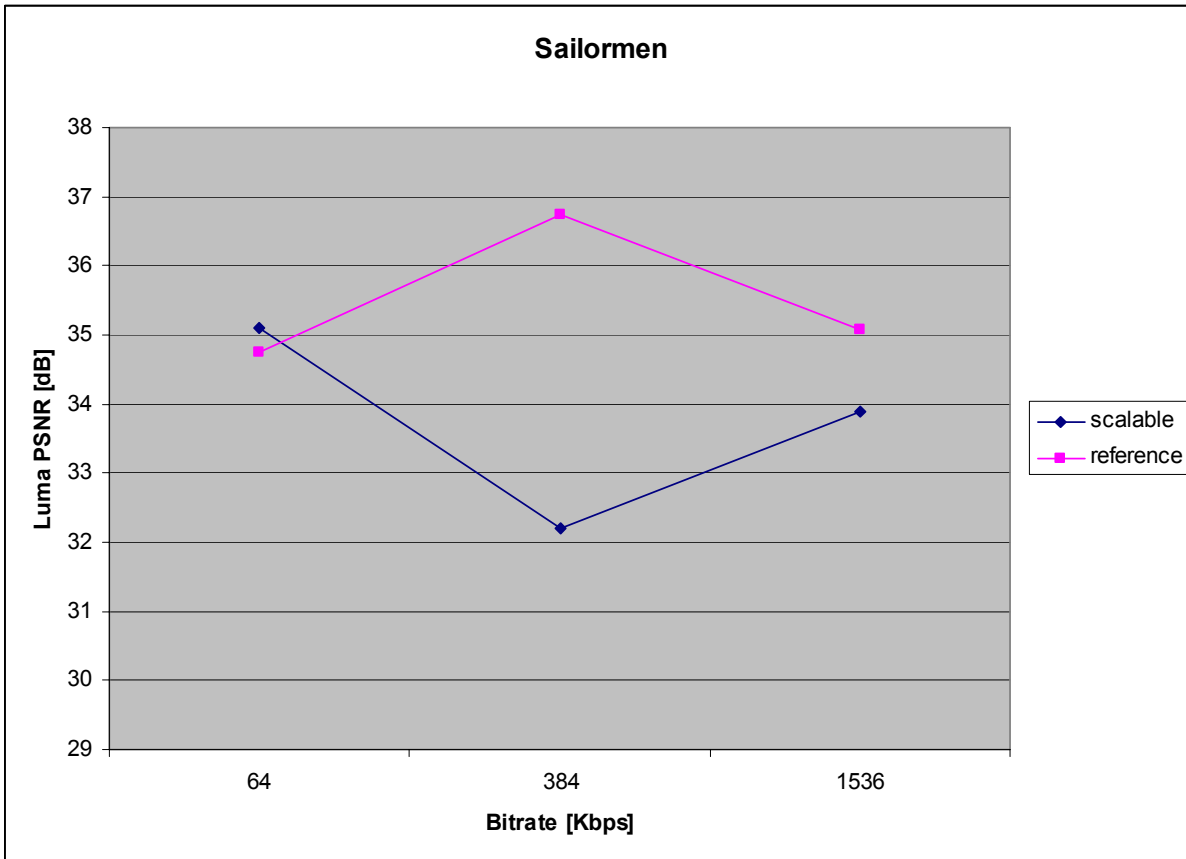


Fig. 6 (cont'd). The results for the test 1a (the first column of the table 1 [6]).

8. Conclusions

Described is a scalable extension of the AVC codec. The basic features of the multi-loop coder structure 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.

The scalable coder exhibits good satisfactory coding performance. Scalable coder complexity is similar to that of the simulcast structure. The major additional operations are:

- spatial interpolation,
- additional mode selection.

More complex is an optional operation that improves coding efficiency:

- additional motion estimation for averages of spatial and temporal reference .

In this contribution, two variants of mixed spatio-temporal prediction are compared. The coding performance is similar for both variants.

The coding performance strongly depends on the bit allocation to the individual layers for the given resolutions. If the bitrates are given the spatio-temporal resolutions may be modified in order to match the better coding performance. Moreover, the scalability overhead is higher for higher number of layers.

The coder works well if the bitrate ratio for the consecutive layers is about 1:3. Here, in the test 1a this ratio was about 1:5 or 1:6 which is far from the optimum working conditions for the structure.

Acknowledgement

The work has been supported by the Polish State Committee for Scientific Research in the years 2003-2005.

References

- [1] D. Wu, Y. Hou, Y. Zhang, "Scalable video coding and transport over broad-band wireless networks," *Proc. of the IEEE*, vol. 89, pp. 6-20, January 2001.
- [2] M. van der Schaar, C.J. Tsai, T. Ebrahimi, Report of ad hoc group on scalable video coding, ISO/IEC JTC1/SC29/WG11 MPEG02/M9076, Dec. 2002.
- [3] M. van der Schaar, C.J. Tsai, T. Ebrahimi, "Report of Ad hoc Group on Scalable Video Coding", ISO/IEC/SC29/WG11/MPEG2003/9358, March 2003.
- [4] K. Ugur, P. Nasiopoulos, "Design Issues and Proposals for H.264 based FGS", ISO/IEC/SC29/WG11/MPEG2003/M9505, March 2003.
- [5] L. Błaszak, M. Domański, S. Maćkowiak, "Spatio-temporal scalability in AVC codecs", ISO/IEC/SC29/WG11/MPEG2003/M9469, March 2003.
- [6] "Call for Evidence on Scalable Video Coding Advances", Doc. ISO/IEC/SC29/WG11/MPEG2003/N5559.
- [7] Ł. Błaszak, M. Domański, A. Łuczak, S. Maćkowiak, "Spatio-temporal scalability in DCT-based hybrid video coders", ISO/IEC JTC1/SC29/WG11 MPEG02/M8672, July 2002.
- [8] M. Domański, A. Łuczak, S. Maćkowiak, "On improving MPEG spatial scalability", *Proc. Int. Conf. Image Proc.*, Vancouver, 2000, vol. 2, pp. 848-851.
- [9] 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.
- [10] S. Kondo, S. Kadono and M. Schlockermann, Proposal of minor changes to multi-frame buffering syntax for improving coding efficiency of B-pictures, ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Document JVT-B057, February 2002.