AVC VIDEO CODERS WITH SPATIAL AND TEMPORAL SCALABILITY

Marek Domański, Łukasz Błaszak, Sławomir Maćkowiak,

Poznań University of Technology, Institute of Electronics and Telecommunication, Poznań, Poland
E-mail: { domanski, lblaszak, smack } @ et.put.poznan.pl

ABSTRACT

The paper describes a scalable extension of the AVC coder. The assumption is to introduce possibly minor modifications of the bitstream semantics and syntax as well as to avoid as much as possible the technologies that are not present in the existing structure of the AVC codec. The coder combines spatial with temporal scalability. The coder consists of two motion-compensated sub-coders that encode a video sequence and produce two bitstreams corresponding to two different levels of spatial and temporal resolution. Each of the sub-coders has its own prediction loop with independent motion estimation. The system employs adaptive interpolation. The interpolation-dependent modes are carefully embedded into the mode hierarchy of the AVC coder thus obtaining the codes that correspond to the mode probabilities.

1. INTRODUCTION

Recently, the JVT Committee has prepared Version 1 of the new video coding standard called AVC [1]. This standard is called also H.264, and defines improved hybrid video coding tools. The main features related to improve coding efficiency are the following: flexible size of rectangular blocks for motion-compensated prediction, advanced intraframe prediction, flexible choice of prediction modes, a multi-frame memory in the motion-compensated predictor. The AVC codec exhibits substantial efficiency improvement as compared to the H.263+ and MPEG-4 natural video codecs.

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. [3,4]

The goal of the paper is to describe a scalable extension of the AVC coder. The assumption is to introduce possibly minor modifications of the bitstream semantics and syntax as well as to avoid as much as possible the technologies that are not present in the existing structure of the AVC codec. Such an approach will limit the implementation costs of scalability.

Currently, two major candidates are considered for future standard scalable video coders [5], i.e.:

- modified hybrid video coder with motion-compensated prediction and block-based transforms,
- wavelet-based video coder with a special emphasis on 3-D wavelet video coder with motion-compensated filter banks.

Considered are also various combinations of the both above mentioned approaches.

Here, the first approach is considered because it does require neither deep modifications of the AVC bitstream syntax nor a major change of the AVC codec structure. In the context of H.26L (the earlier version of the AVC coder), similar approach was already exploited as described in [2]. Nevertheless the approach from [2] has employed a different coder structure with common motion estimation which resulted in worse motion compensation.

2. CODER STRUCTURE

This paper deals with coders that can combine spatial and temporal scalability. In such a case, the base layer represents a video sequence with reduced both temporal and spatial resolutions. Such a combination is very practical as it allows low bitrate base layer with reasonable spatial resolution, e.g. CIF/SIF for standard television input.

Our scalable coder adopts the structure already proposed for MPEG-2 and H.263 codecs [6-8]. It consists of two motion-compensated sub-coders (Fig. 1) that produce two bitstreams corresponding to two different levels of spatial and temporal resolution. Each of the sub-coders has its own prediction loop with independent motion estimation.

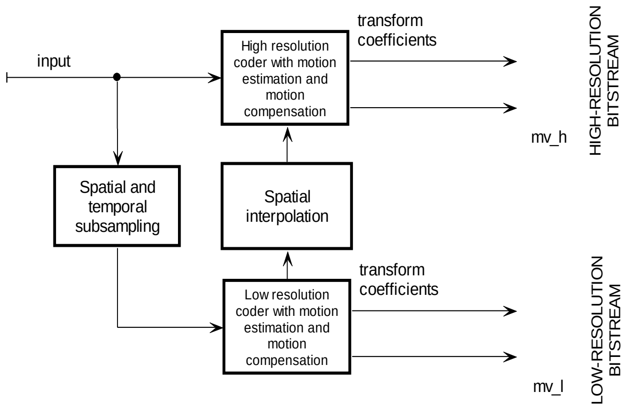
input
High resolution coder with motion estimation and motion compensation
transform coefficients
mv_h
HIGH-RESOLUTION BITSTREAM
Spatial and temporal subsampling
Spatial interpolation
Low resolution coder with motion estimation and motion compensation
transform coefficients
mv_l
LOW-RESOLUTION BITSTREAM

Fig. 1. General structure of the scalable coder considered. *mv_l* and *mv_h* denote motion vectors from the low-resolution and the high-resolution layer, respectively.

The low-resolution sub-coder is implemented as a standard motion-compensated hybrid AVC coder that produces a bitstream with fully standard AVC syntax. The high-resolution sub-coder is a modified AVC coder that is able to exploit the interpolated macroblocks from the decoded base-layer bitstream. These interpolated macroblocks are used as reference macroblocks for prediction whenever they provide lower cost. Other additional reference macroblocks are created by averaging the reference of temporal prediction and the interpolated macroblock.

3. SPATIO-TEMPORAL DECOMPOSITION

Good performance of spatio-temporal down- and upsampling is critical for good performance of the whole codec.

Spatial decimation includes spatial lowpass filtering that prevents spatial aliasing in the base-layer low-resolution sequence. The choice of the filter trades off between high aliasing attenuation and short temporal response. The results of experimental comparisons prove the importance of the careful choice of the decimation-interpolation scheme.

The system considered employs edge-adaptive bi-cubic interpolation as described in [9]. The technique is applicable to both luminance and chrominance.

The technique of edge-adaptive interpolation is an extension of the standard non-adaptive bi-cubic separable interpolation that can be described as follows. The two-dimensional interpolation is performed in two steps: horizontal and vertical. Let $f(x)$ is the value to be interpolated, and the nearest available values are located at coordinates x_k (left) and x_{k+1} (right). Let

$$s = x - x_k\ ,\ 1 - s = x_{k+1} - x\ ,\ \text{where } 0 \leq s \leq 1.$$

There is

$$f(x) = f(x_{k-1})(-s^3 + 2\,s^2 - s)/2 + f(x_k)(3s^3 - 5s^2 + 2)/2 + f(x_{k+1})(-3s^3 + 4s^2 + s)/2 + f(x_{k+2})(s^3 - s^2)/2,$$

where x_{k-1}, x_k, x_{k+1} and x_{k+2} are the positions of four neighboring known pixels.

In the edge-adaptive scheme, a modified value s' is used instead of s.

$$s' = s - kAs(s - 1),$$

where k is a positive parameter that controls the intensity of warping and A is a function of asymmetry of the data in the neighborhood of x:

$$A = (\ |f(x_{k+1}) - f(x_{k-1})| - |\ f(x_{k+2}) - f(x_k)|\ )/(L - 1),$$

where L = 256 for 8-bit sample representation. In the experiments, it was k = 3.05.

In the simplest case, temporal downsampling is performed via frame skipping. In particular, B-frame skipping constitutes very efficient and robust downsampling scheme (Fig. 2).

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.

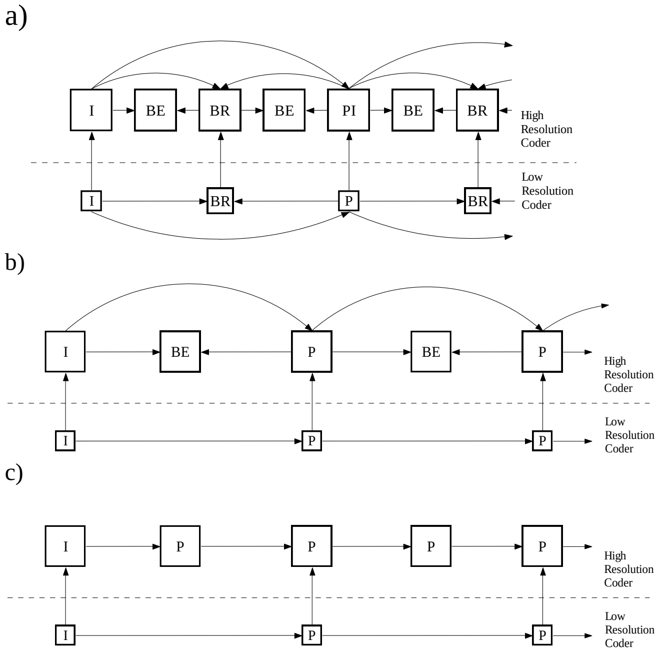

Fig. 2. Exemplary structures of low-resolution and high-resolution video sequences with temporal subsampling by factor 2.

4. REFRENCE FRAMES

In the enhancement layer, the coding scheme takes advantage of two additional reference frames:

- the frame interpolated from the decoded current base-layer low-resolution frame,
- an average of the latter and the last temporal reference frame.

One can use even more reference frames obtained as combinations of the interpolated frame and various temporal reference. Nevertheless those possibilities have been not exploited in the experiments reported in a subsequent paragraph.

Moreover, for the latter above mentioned reference frame, 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. This option was used in the experiments reported further.

Application of these additional reference frames does not require bitstream syntax modifications and just minor modifications of the semantics for the reference frame variables.

5. PREDICTION MODE SELECTION

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.

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

Frame type	Prediction modes
Intra (I)	1. Spatial interpolation from base layer (16×16 block size). 2. All standard intra prediction modes.
Inter (P)	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)	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.

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.

In order to test the coding performance of the scalable AVC codec, a series of experiments have been performed with (352×288)-pixel sequences.

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

- CABAC coder,
- ¼-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.

Table 2. Coding efficiency comparison for scalable, nonscalable and simulcast coding

Test sequence	*Bus*		*Cheer*		*Football*		*Fun*		*Basket*	
Original frame rate [fps]	30		30		30		25		25	
QP_I=10, QP_P=11, QP_B=12										
	PSNR [dB]	Bitrate [kbps]	PSNR [dB]	Bitrate [kbps]	PSNR [dB]	Bitrate [kbps]	PSNR [dB]	Bitrate [kbps]	PSNR [dB]	Bitrate [kbps]
Base layer	37.17	502.11	38.60	279.35	37.72	751.80	39.76	354.37	37.17	502.11
Enhancement layer	38.06	2215.86	39.06	1533.10	38.56	3030.11	40.75	1388.49	38.06	2215.86
Whole scalable codec	38.06	2717.97	39.06	1812.45	38.56	3781.91	40.75	1742.86	38.06	2717.97
Simulcast	38.02	3039.57	39.05	1939.98	38.54	4384.43	40.75	2081.16	38.02	3039.57
Nonscalable codec	38.02	2537.46	39.05	1660.63	38.54	3632.63	40.75	1726.79	38.02	2537.46
QP_I=15, QP_P=16, QP_B=17										
	PSNR [dB]	Bitrate [kbps]	PSNR [dB]	Bitrate [kbps]	PSNR [dB]	Bitrate [kbps]	PSNR [dB]	Bitrate [kbps]	PSNR [dB]	Bitrate [kbps]
Base layer	32.96	286.13	34.65	151.94	33.51	457.17	36.20	200.74	32.96	286.13
Enhancement layer	34.08	1189.34	35.14	816.94	34.51	1702.18	37.31	759.71	34.08	1189.34
Whole scalable codec	34.08	1475.47	35.14	968.88	34.51	2159.35	37.31	960.45	34.08	1475.47
Simulcast	34.08	1668.77	35.16	1036.59	34.55	2559.36	37.45	1165.98	34.08	1668.77
Nonscalable codec	34.08	1382.64	35.16	884.65	34.55	2102.19	37.45	965.24	34.08	1382.64
QP_I=20, QP_P=21, QP_B=22										
	PSNR [dB]	Bitrate [kbps]	PSNR [dB]	Bitrate [kbps]	PSNR [dB]	Bitrate [kbps]	PSNR [dB]	Bitrate [kbps]	PSNR [dB]	Bitrate [kbps]
Base layer	29.06	156.33	31.03	77.85	29.65	261.20	33.12	109.05	29.06	156.33
Enhancement layer	30.28	645.36	31.52	434.19	30.76	955.72	34.17	420.95	30.28	645.36
Whole scalable codec	30.28	801.69	31.52	512.04	30.76	1216.92	34.17	530.00	30.28	801.69
Simulcast	30.33	913.55	31.63	544.96	30.89	1466.69	34.39	651.76	30.33	913.55
Nonscalable codec	30.33	757.22	31.63	467.11	30.89	1205.49	34.39	542.71	30.33	757.22

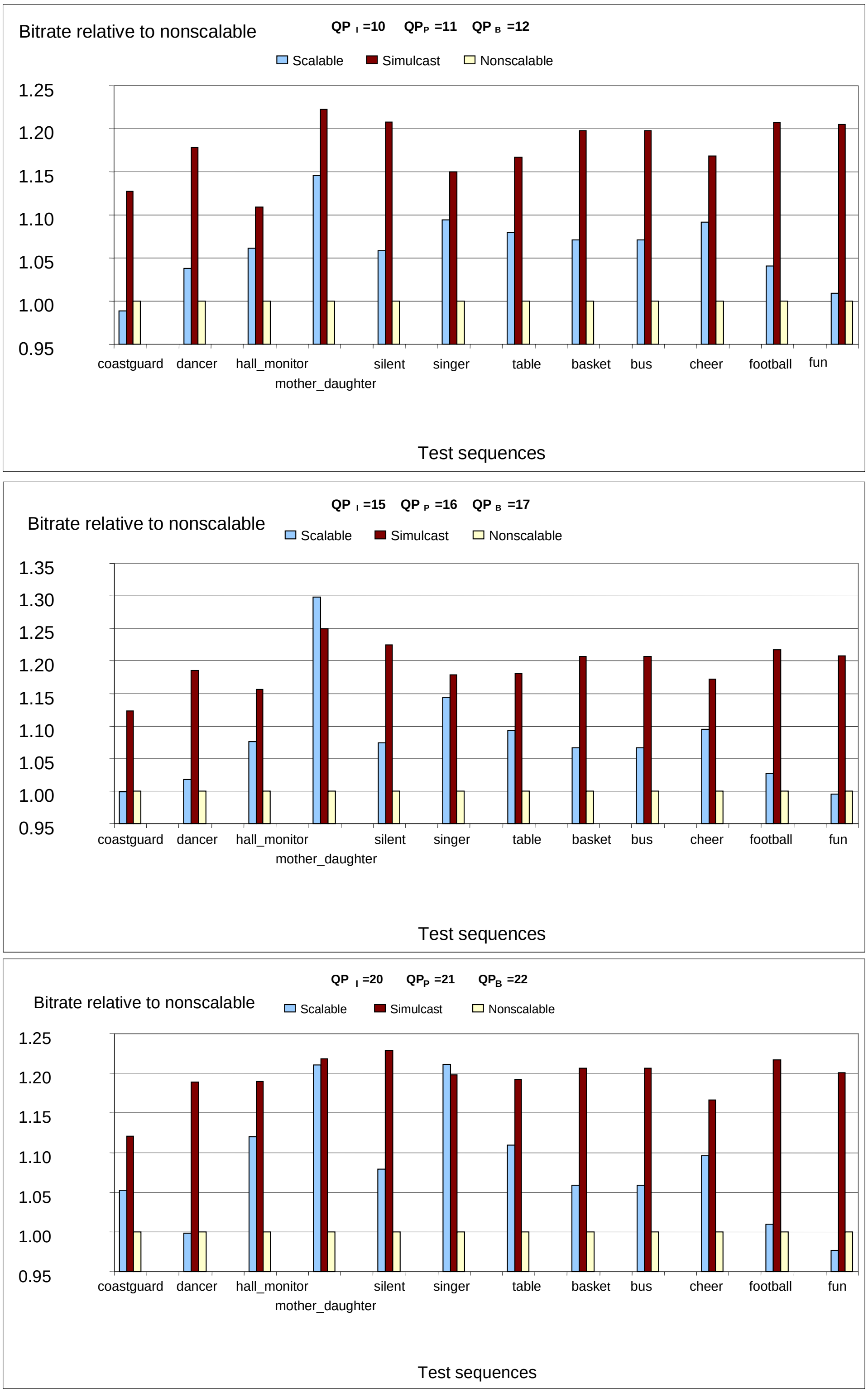

Fig. 3. Approximate bitrate comparison for scalable, nonscalable (single-layer) and simulcast coding.

In order to compare the scalable codec with the nonsacalable reference AVC codec as well as with the simulcast pair of nonscalable 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, nonscalable 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*), nonscalable coder and simulcast coding (Table 2 and Fig. 3).

For such test conditions, the approximate bitrate overhead due to scalability was between -1% and 30% of the bitrate for the nonscalable (single-layer) codec (Fig. 3). 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.

7. CONCLUSIONS

For the two-layer system with spatio-temporal scalability, the bitrate overhead due to scalability varies between -1% and 30% depending on sequence content and bitrate allocation (Fig.3).

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.

In the paper, described is an extension of AVC coder structure. The major features of the presented solution are:

- mixed spatio-temporal scalability,
- independent motion estimation for each motion-compensation loop, i.e. for each spatio-temporal resolution layer,
- adaptive decimation and interpolation.

These above features are also the reasons for good performance of the whole coder.

ACKNOWLEDGEMENT

The work has been partially supported by the IST Programme of the EU under contract number IST-2000-26467 (MASCOT).

REFERENCES

[1] ISO/IEC/SC29/WG11/MPEG02/N4920, ISO/IEC 14496-10 AVC │ ITU-T Rec. H.264, Text of Final Committee Draft of Joint Video Specification, Klagenfurt, July 2002.

[2] Y. He, R. Yan, F. Wu, S. Li, H.26L-based fine granularity scalable video coding, ISO/IEC JTC1/SC29/ WG11 MPEG02/M7788, Dec. 2001.

[3] 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.

[4] 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.

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

[6] M. Domański, S. Maćkowiak, “On improving MPEG spatial scalability”, in *Proc. Int. Conf. Image Proc.,* vol. 2, pp. 848-851, 2000.

[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] S. Maćkowiak, “Scalable Coding of Digital Video”, Doctoral dissertation, Poznań University of Technology, Poznań 2002.

[9] G. Ramponi, Warped distance for space-variant linear image interpolation, *IEEE Transactions on Image Processing*, vol. 8, pp. 629-639, May 1999.