Extended AVC/H.264 Video Codecs with Mixed Scalability

Łukasz Błaszak, Marek Domański, Sławomir Maćkowiak
Institute of Electronics and Telecommunication
Poznań University of Technology
Poznań, Poland
e-mail: [lblaszak, domanski, smack] @et.put.poznan.pl

Abstract— **The paper describes a scalable extension of the AVC/H.264 coder. The proposed coder combines spatial and temporal scalability with FGS (Fine Granularity Scalability). The solution proposed introduces minor modifications of the bitstream semantics and syntax. Decimation and interpolation are the only functions that correspond to codec building blocks that are not present in the existing structure of the AVC codec. The coder consists of two independently motion-compensated sub-coders that encode a video sequence and produce two bitstreams corresponding to two different levels of spatial and temporal resolution. The system employs adaptive interpolation as well as luminance-assisted interpolation of chrominance. The functionality of FGS is related to some drift in the enhancement layer. This drift can be limited by excluding temporal prediction in some enhancement layer frames.**

Keywords: ***hybrid video codecs; spatio-temporal scalability;AVC***

I. INTRODUCTION

MPEG-4 standard itself contains various encoding methods, called "subsets" or "layers", and the H.264 [1] is a latest standardized layer in the MPEG-4. Similar to former video standards H.264 is also based on the hybrid coding structure. The characteristic features of AVC/H.264 coder are:

- The encoder uses block-based 4×4 and 2×2 integer transforms; In contrary to existing encoders such as H.263,
- MPEG-4, it has very flexible size of rectangular blocks for motion-compensated prediction;
- It uses multi-frame memory in motion-compensated prediction.
- The coding performance is much better then standard encoders', such as H.263, MPEG-2.

The current version of AVC encoder does not support scalability. Due to the fact that such functionality is very important nowadays, it is vital to include it into this new advanced codec.

The paper describes a scalable extension of the AVC encoder. The technique provides a combination of spatial, temporal and FGS scalability. The authors' objective was to introduce only minor modifications to the AVC bitstream as well as to the codec structure. Another authors' objective was to design codec with possibly low computational cost being not essentially higher than that of simulcast coder pair, i.e. two independent AVC coders for two resolutions.

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. In the references [14-20], described are also other solutions based on modified hybrid video codecs with motion-compensated prediction and block-based transforms. These solutions usually adopt deeper changes of the codec structures. Even more dramatic change of coding technology is related to wavelet-based video codecs that exhibit flexible scalability. Recently, 3D wavelet video coders with motion-compensated filter banks [6-13,21] have gained a lot of attention Nevertheless, their application would need substantial change of coding technology.

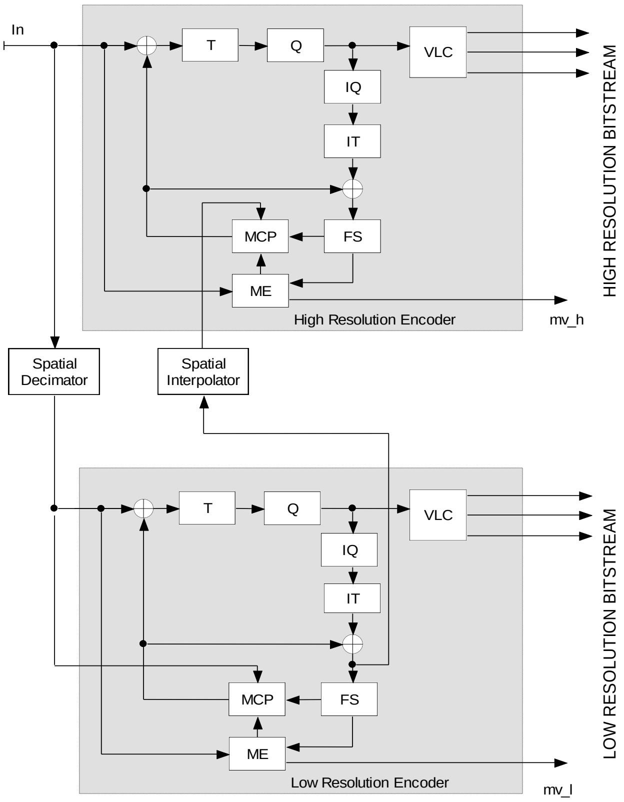
II. CODER STRUCTURE

The paper describes an additional feature in the AVC (H.264 encoder), which is spatio-temporal scalability with fine granularity scalability. It provides a possibility to produce one bitstream which represents an encoded sequence with two different spatial and temporal resolutions at the same time. The very important thing is that produced bitstream is smaller then sum of two separately encoded sequences.

The bitstream at the output of the scalable encoder consists of two layers. One is a base layer that is decimated in time and in space. This layer is fully compatible with standard H.264 bitstream syntax, so this layer can be decoded by any standard AVC decoder. This feature of scalable encoder bitstream syntax is very important, because new functionality such as scalability should not influence on behavior of decoders that do not recognize additional data in a bitstream. The second layer in bitstream is an enhancement layer. It consists of data which are needed to decode a full-quality video sequence. In order to decode the enhancement layer video sequence, the decoded base layer sequence is also needed.

In fact, our encoder consists of two motion-compensated sub-coders (Fig. 1). Each of the sub-coders has its own prediction loop with independent motion estimation and compensation. Data partitioning is used in order to obtain the FGS functionality. Further detailed considerations and experiments deal with data partitioning in the high-resolution enhancement layer only. For the horizontal, vertical and temporal subsampling factors of 2, the range of bitrate matching due to FGS extends mostly from about 30% to 100% of the total bitrate for a scalable coder.

Each encoder produces one layer of a bitstream. The structure of video sequence in such a bitstream may take different schemas. An exemplary one is shown at the Fig. 2.



T	Transform (e.g. DCT)
Q	Quantization
IQ	Dequantization
IT	Inverse transform
FS	Frame memory
MCP	Motion-compensated predictor
ME	Motion estimator

Figure 1. The structure of the encoder (temporal subsampling is not included in this figure). *VLC – variable-length coder*. *mv_l* and *mv_h* denote motion vectors from the low-resolution and the high-resolution layer, respectively.

Encoder of base layer is not modified H.264 encoder, but as an input it takes a sequence decimated in time and in space. For this purpose 12th order zero-phase filters with flat passband attenuation characteristics and passband cutoff frequency of about 0.4 of the Nyquist frequency has been used:

$$h(n) = [\ 0.038546219, 0.016179909, -0.057469217, -0.070531366, 0.071806408, 0.297291427, 0.408353238, 0.297291427, 0.071806408, -0.070531366, -0.057469217, 0.016179909, 0.038546219].$$

It is very important to use a proper decimation process, because it has strong influence on encoding performance. Several FIR filters have been tested by authors in order to find currently used one. Temporal downsampling is performed via frame skipping. In particular, B-frame skipping constitutes very efficient and robust downsampling scheme.

Encoder of enhancement layer is based on H.264 encoder. It takes as an input additional video sequence, which is decoded base layer sequence. Avery frame of this sequence is being interpolated to the resolution of frames in enhancement layer video sequence. The interpolation process, like decimation, is also very important, because of strong influence on performance of encoding enhancement layer. In our test model encoder we used a G. Ramponi technique [21] which is edge adaptive interpolation. As a base function the bicubic one has been chosen,

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

where x_k, x_{k+1}, x_{k+2}, x_{k-1} are neighbors of the x, and s is a distance between the first neighbor of x and x. The value s is modified:

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

where

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

Parameter k has been experimentally estimated and set to value 3.05, L =256 for 8 bit pixel representation.

Enhancement layer encoder has extended list of prediction modes. Those modes correspond to additional information provided into this encoder from base layer decoder. Those modes are described later in this paper.

Efficient scheme to encode the enhancement-layer motion vectors is described in the full paper. In principle, it is a modified AVC scheme that exploits encoding of the residuals from median prediction of the motion vectors components. In some situations, the prediction is enhanced by the information about the corresponding motion vector from the low-resolution base layer.

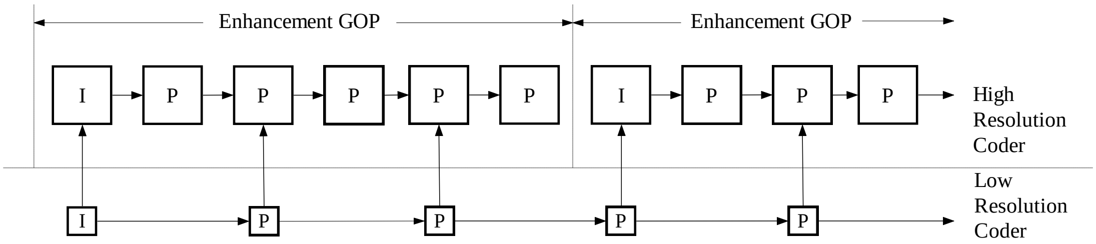


Figure 2. Exemplary structure of a video sequence: No B-frames and no GOP structure in the base layer. In the enhancement layer, the I-frames are encoded with respect to the interpolated I- or P-frames from the base layer.

III. FINE GRANULARITY SCALABILITY

Fine granularity scalability (FGS) is obtained via data partitioning in the UVLC (exp-Golomb) coding mode. Drift propagation is limited by insertion of I-frames into the enhancement layer (Figure 2). Such additional enhancement-layer I-frames are encoded using less numbers of bits than single-layer I-frames. It is because the bitstream syntax of these frames is that of P-frames but with no motion vectors and with the interpolated base-layer frames used as reference frames.

IV. PREDICTION MODE SELECTION

Scalable AVC encoder's prediction mode list has been extended. New modes correspond to the additional information available in enhancement layer extracted from base layer. There are mainly two additional modes: First one takes as a prediction an interpolated block from base layer; second one takes as a prediction block which is an average of interpolated block from base layer and prediction block from previous frame. The encoder is looking for such a block in previous frame which after averaging with interpolated block gives the smallest prediction error.

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 I.

TABLE I. 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.

When the mode with only interpolation is chosen, there is no motion vectors we need to send to decoder. So the better the interpolation process is the often the interpolation mode is chosen. And because of no motion vectors this mode is very efficient. In the averaging mode we have to consider a cost of the mode. Mainly two things have strong influence on this cost: count of bits needed for: motion vectors and prediction error.

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.

V. CODING PERFORMANCE AND CONCLUSIONS

Coding performance is bounded by the performance of two extreme boundary systems:

- simulcast, i.e. the two layers encoded independently,
- single-layer coding.

For the two-layer system with spatio-temporal scalability, the bitrate overhead due to scalability varies between 9% and 25% depending on sequence content and bitrate allocation (Table 2). The full paper includes detailed experimental results including those for FGS. The results have been obtained using a test model upgraded from the AVC ver. 2.1 model. The CABAC-based and UVLC-based coder were used in the experiments but FGS was implemented for the UVLC (exp-Golomb) version only.

Coding efficiency may be improved by enhancing the decimation and interpolation as mentioned before. The results from Table 2 have been obtained without using these options.

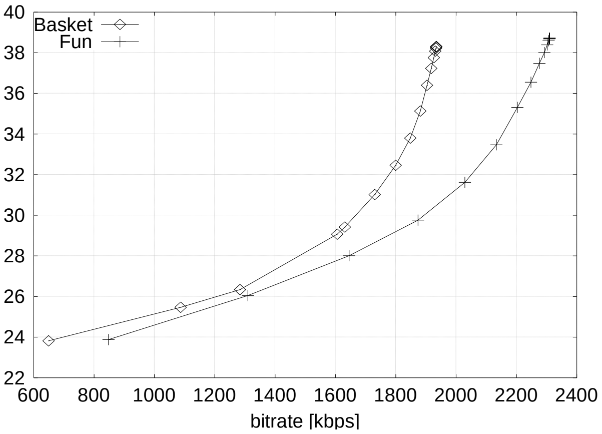


Figure 3. Rate-distortion curves for FGS in the extended AVC codec: test sequences *Fun* and *Basket.*

The range of bitrate matching due to FGS extends mostly from about 30% to 100% of the total bitrate for a scalable coder (Figure 3)

In the paper, described is a generic coder structure for motion-compensated fine-granularity scalability. The major differences with respect to the proposal from [2] 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.

TABLE II. CODING PERFORMANCE FOR TWO-LAYER SYSTEM WITH SPATIAL AND TEMPORAL SCALABILITY (RESULTS ARE LISTED FOR VARIOUS QUANTIZATION IN THE BASE LAYER).

12 order interpo-lation filter	***Cheer* sequence**		***Football* sequence**	
	Bitstream [kbit/s]	***Luminance PSNR [dB]***	***Bitstream [kbit/s]***	***Luminance PSNR [dB]***
Base layer	353.22	33.68	148.00	36.59
Enhance-ment layer **QI=15** **QP=16** **QB=17**	1249.29	34.69	550.64	37.72
Non-scalable **QI=15** **QP=16** **QB=17**	1555.74	34.73	680.81	37.80
	Total bitrate	***Overhead [%]***	***Total bitrate***	***Overhead [%]***
Scalable	1602.51	3.0	698.64	2.6
Simulcast	1908.96	22.7	828.81	21.7

ACKNOWLEDGEMENT

This work has been supported by the Polish National Committee for Scientific Research under a research project in years 2003-2005.

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 Proc. , 2001.

[6] K. Shen and E. Delp, “Wavelet based rate scalable video compression, IEEE Trans. Circ. Syst. Video Techn.” vol. 9, pp. 109-122, February 1999.

[7] B.-J. Kim, Z. Xiong and W. Pearlman, “Low bit-rate scalable video coding with 3-D set partitioning in hierarchical trees (3-D SPIHT),” IEEE Trans. Circ. Syst. Video Techn., v. 10, pp. 1374-1387, Dec. 2000.

[8] J.-Y. Lee, H.-S. Oh and S.-J. Ko, “Motion-compensated layered video coding for playback scalability,” IEEE Trans. Circ. Syst. Video Techn., vol. 11, pp. 619-628, May 2001.

[9] J.-R. Ohm, "Three-dimensional subband coding with motion compensation," IEEE Trans. Image Proc., vol. 3, pp. 559-571, Sept. 1994.

[10] V. Bottreau, M. Benetiere, B. Felts and B. Pesquet-Popescu, “A fully scalable 3D subband video codec”, Proc. Int. Conf. Image Processing, ICIP’2001, vol. II, pp. 1017-1020, Thessaloniki, October 2001.

[11] Z. Zhang, G. Liu, Y. Yang, “High performance full scalable video compression with embedded multiresolution MC-3D-SPIHT”, Proc. Int. Conf. Image Proc., vol. III, pp. 721-734, Rochester, NY, Sept. 2002.

[12] A. Secker, D. Taubman, “Highly scalable video compression using a lifting-based 3D wavelet transform with deformable mesh motion compensation”, in Proc. Int. Conf. Image Proc., vol. III, pp. 749-752, 2002.

[13] Y. Andreopoulos, et al, “Scalable wavelet video-coding with in-band prediction – implementation and experimental results”, in Proc. Int. Conf. Image Proc., vol. III, pp. 729-732, 2002.

[14] W. Li, “Overview of fine granularity scalability in MPEG-4 video standard,” IEEE Trans. Circ. Syst. Video Techn., vol. 11, pp. 301-317, March 2001.

[15] S. Regunathan, R. Zhang and K. Rose, “Scalable video coding with robust mode selection,” Signal Processing: Image Communication, vol. 16, pp. 725-732, May 2001.

[16] K. Rose and S. Regunathan, “Toward optimality in scalable predictive coding,” IEEE Trans. Circ. Syst. Video Techn., vol. 11, pp. 965-976, July 2001.

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

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

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

[20] 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.

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