COMBINED SCALABILITY SUPPORT FOR THE SCALABLE EXTENSION OF H.264/AVC

Heiko Schwarz, Detlev Marpe, Thomas Schierl, and Thomas Wiegand

Fraunhofer Institute for Telecommunications – Heinrich Hertz Institute, Image Processing Department
Einsteinufer 37, D-10587 Berlin, Germany
{hschwarz,marpe,schierl,wiegand}@hhi.fraunhofer.de

ABSTRACT

The scalability extension of H.264/AVC [1] uses a layered approach for providing spatial, temporal, and SNR scalability. Due to this concept, only a restricted set of spatio-temporal-SNR points can be extracted and decoded from a global scalable bit-stream, and this set of points is defined by the chosen encoder configuration. In this paper, we present a new approach for providing flexible combined spatial, temporal, and SNR scalability. The increased flexibility is achieved by introducing NAL units that represent a refinement signal for a picture in a coarse-to-fine-description and can be truncated at any arbitrary point. The simulation results show that this approach is capable of providing flexible combined scalability while the coding efficiency is only slightly worse than that of the layered approach.

1. INTRODUCTION

The scalable extension of H.264/AVC as proposed in [1][2] has been chosen to be the starting point of MPEG's Scalable Video Coding (SVC) project in October 2004. In January 2005, MPEG and the Video Coding Experts Group (VCEG) of the ITU-T agreed to jointly finalize the SVC project as an Amendment of their H.264/AVC standard [3][4], and the scalable extension of H.264/AVC was selected as the first Working Draft [5].

The basic design idea of the scalable H.264/AVC extension is to extend the hybrid video coding approach of H.264/AVC [3][4] towards motion-compensated temporal filtering (MCTF) by using a lifting framework. Because lifting is invertible, any motion compensation (MC) technique can be incorporated into the prediction and update steps of the filter bank. By using the highly efficient motion model of H.264/AVC [3][4] in conjunction with a block-adaptive switching between the Haar and the 5/3 spline wavelet, both the prediction and the update step are similar to MC techniques in the generalized B slices of H.264/AVC.

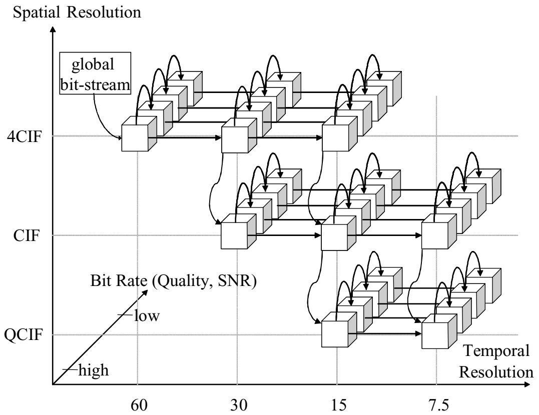

Fig. 1: Concept of flexible combined scalability.

Furthermore, the open-loop structure of MCTF offers the possibility to efficiently incorporate spatial, temporal, and SNR scalability. In this paper, we present the combination of these which is referred to as combined scalability. Fig. 1 illustrates the concept, where each box corresponds to a possible decoded video signal. The video signal that is represented in the global scalable bit-stream is characterized by temporal and spatial resolution as well as a maximum bit rate, or equivalently SNR at a particular resolution. Within an admissible range of these 3 parameters, a point can be selected at which the bit-stream can be extracted and decoded. This extracted bit-stream again contains a variety of temporal and spatial resolutions and bit rates of the decoded video signal. It should be noted that in practical scenarios there are constraints on the contained resolutions.

2. MCTF EXTENSION OF H.264/AVC

H.264/AVC is a hybrid video codec specifying for macroblocks either motion-compensated prediction or intra prediction. Both predictions are followed by residual coding [3][4]. When lifting is used to implement an MCTF, the prediction and update steps are separate mechanisms. Hence, we use the MC of H.264/AVC for the prediction step and a similar technique for the update step which is derived from the prediction step. Thus the update

0-7803-9332-5/05/$20.00 ©2005 IEEE

step also consists of block-based H.264/AVC MC, but with a bit-depth expansion by 1 compared to the prediction step. The algorithm for the derivation of the motion vectors and the reference picture indices for the update step is given in [1][5].

When MC does not work, e.g. for scene cuts or uncovered background, the incorporation of intra coding modes increases coding efficiency. For the intra macroblock, the corresponding prediction or update step is skipped and coded using the intra coding tools of H.264/AVC. Note, that these intra samples are set to zero before they are used for MC in the update steps.

The temporal coding structure of MCTF is changed relative to hybrid video coding in that not only high-pass pictures H_k (prediction residuals) are resulting from the prediction step but also low-pass pictures L_k are resulting from the update step. Typically, a group of N_0 input pictures is partitioned into two sets of pictures with one set containing N_A ($0 < N_A < N_0$) input pictures and the other set containing $N_B = N_0 - N_A$ input pictures. These two sets are used for a two-channel MCTF decomposition, where the pictures of the first set are spatially shift aligned with the low-pass pictures L_k, and the pictures of the second set are spatially shift aligned with the high-pass pictures H_k. In general, the two-channel decomposition is iteratively applied to the set the low-pass pictures until a single low-pass picture is obtained or a given number of decomposition stages is performed.

3. SCALABILITY EXTENSION OF H.264/AVC

In this work, we refer to scalability as a functionality that allows the removal of parts of the bit-stream while achieving a reasonable coding efficiency of the decoded video at reduced temporal, SNR, or spatial resolution. A bit-stream consists of a base layer and one or more enhancement layers that are nested.

3.1. Temporal Scalability

The temporal decomposition as described in Sec. 2 permits temporal scalability in a similar way as in hybrid video coding. The scalability is achieved by removing those bit-stream parts that correspond to pictures that are not reference pictures for the prediction of the retained pictures. However, it should be noted that for resolutions lower than the maximum temporal resolution at a particular spatial resolution, the decoded pictures are low-pass (L) pictures that are constructed from typically 2 or more other pictures of the higher temporal resolution.

3.2. Spatial Scalability

We consider spatial scalable coding of video at multiple resolutions (e.g. QCIF, CIF, and 4CIF) with a factor of 2 in horizontal and vertical resolution. The video is generally represented using an oversampled pyramid. The following switchable inter-layer prediction techniques turned out to provide gains and are used in the codec:

1. Prediction of a macroblock using the up-sampled lower resolution signal
2. Prediction of motion vectors using the up-sampled lower resolution motion vectors
3. Prediction of the residual signal using the up-sampled residual signal of the lower resolution layer

For details on these techniques please refer to [1][2][5].

3.3. New SNR Scalability Approach

For the SNR base layer, H.264/AVC-compatible transform coding is used. The high-pass pictures contain intra or residual macroblocks as in hybrid video coding. For the residual macroblocks, the coding as in H.264/AVC including transformation and quantization is employed. The intra macroblocks are coded using the intra coding modes of H.264/AVC. For each macroblock, the residual blocks are transmitted together with the macroblocks modes, intra prediction modes, reference picture indices, and motion vectors using the B or P slice syntax of H.264/AVC. Low-pass pictures are either coded independently of each other as H.264/AVC intra pictures or are inter coded as H.264/AVC inter pictures.

On top of the SNR base layer, the SNR enhancement layer is coded. In [1] a simple extension is described in which the quantization error between the SNR base layer and the original subband pictures is transformed, quantized, and coded exactly using the same methods as for the base layer but with a finer quantization step size. With this approach only coarse grains of scalable SNR layers can be efficiently represented such as factors of 2 in bit rate. We have therefore investigated an approach, which enables an efficient representation of finer grains of SNR scalability.

In that scheme, low- and high-pass pictures for a spatial resolution are generally represented by a base layer that includes the motion data and is coded using the H.264/AVC syntax (cp. [1]) and zero or more progressive refinement layers. Usually the base layer corresponds to a minimally acceptable reconstruction quality and this basic quality can be improved in a fine granular way by truncating the enhancement layer NAL units at any arbitrary point. Each enhancement layer packet contains a refinement signal that corresponds to a bisection of the quantization step size. In contrast to the SNR scalability scheme that was proposed in [1], the refinement signals for the low- and high-pass pictures are directly coded in the transform coefficient domain. Thus, at the decoder side, the inverse transform has to be performed only once for each transform block of a subband picture.

In order to provide SNR enhancement layer NAL units that can be truncated at any arbitrary point (progressive refinement NAL units), the coding order of the transform coefficient levels has been modified. For transmitting the refinement representations of the transform coefficients, we re-use the CABAC contexts that are specified in H.264/AVC.

For each refinement representation of a subband picture, which corresponds to a bisection of the quantization step size and is transmitted in a separate NAL unit, the coding process for the transform coefficient refinement levels is divided into 3 scans as follows.

1. In the first scan, the refinement levels of all transform coefficients with the following properties are coded:
    - The transform coefficient levels that have been coded in the base layer representation and all subordinate enhancement layer representations are equal to zero (non-significant transform coefficient).
    - The transform coefficient is located inside a transform block that at least includes one transform coefficient, for which a transform coefficient level not equal to zero has been transmitted in the base layer representation or any subordinate enhancement layer representation (significant transform coefficient block).
2. In the second scan, the refinement levels of all transform coefficients with the following properties are coded:
    - A transform coefficient level not equal to zero has been coded in the base layer or any previous enhancement layer representation (significant transform coefficient).
3. Finally, in the third scan, all remaining refinement levels are coded. The corresponding transform coefficients have the following properties:
    - The transform coefficient levels that have been coded in the base layer representation and all subordinate enhancement layer representations are equal to zero (non-significant transform coefficient).
    - The transform coefficient is located inside a transform block that does not include any transform coefficient, for which a transform coefficient level not equal to zero has been transmitted in the base layer representation or any subordinate enhancement layer representation (non-significant transform coefficient block).

3.4. Combined Scalability

The SNR scalability concept described above is used to enable flexible combined scalability. The general approach is depicted in Fig. 2. The input video (e.g. 4CIF) is first decimated to obtain lower spatial resolution video (e.g. CIF and QCIF). Each of these spatial layers is decomposed by an MCTF as described in Sec. 2 followed by blockwise transform and H.264/AVC-based quantization (base layer coding) followed by FGS coding as described in Sec. 3.3. Starting with the lowest spatial resolution (e.g. QCIF), the syntax elements for the MCTF and the other steps can be passed as predictors to the next higher resolution layer (e.g. CIF) and be used to reduce bit rate as described in Sec. 3.2. Then, the same can be repeated for the next higher layer (e.g. 4CIF).

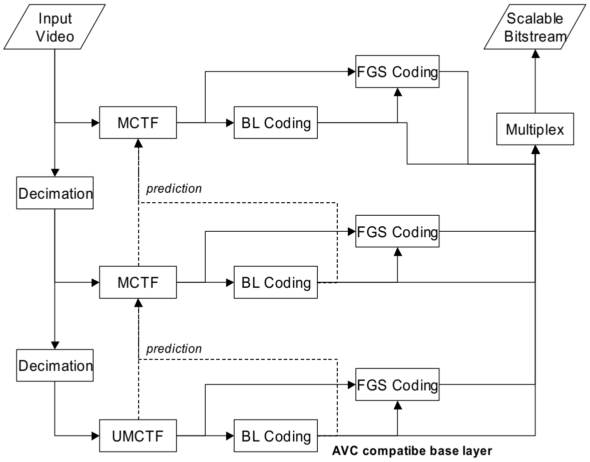

Fig. 2: General approach for flexible combined scalability.

Within the combined scalable bit-stream, the FGS NAL units can be truncated at any arbitrary point. For any spatio-temporal resolution, a minimum bit rate, which represents the corresponding spatial base layer representation (including the base layer representations of the lower-resolution layers), must be transmitted. These bit rates can be adjusted in a way that the corresponding reconstructions represent the minimally acceptable video quality. Above the minimum bit rate for a spatio-temporal resolution, any bit rate can be extracted by truncating the FGS NAL units of the corresponding spatio-temporal layer and all lower resolution layers in a suitable way.

4. EXPERIMENTAL RESULTS

Coding results are presented for the sequence Crew in Fig. 3 and the sequence Soccer in Fig. 4. For both sequences, the spatial and temporal resolutions and corresponding bit rates are shown in Tab. 1.

The combined scalability case as proposed in this paper represents the video signal for all bit rates and resolutions of Tab. 1. Note that all bit rates are nested in a way that typically for a particular bit rate in Tab. 1, all bit-streams to the left and/or above relative to this bit-stream in Tab. 1 are embedded and can be extracted.

Tab. 1: Spatial/temporal resolutions and bit rates

Format	Bit rates (kbit/sec)				
QCIF 15Hz	***96***	112	128	160	***192***
CIF 7.5Hz	192	224	256	320	384
CIF 15Hz	256	320	384	448	512
CIF 30Hz	***384***	448	512	640	***768***
4CIF 15Hz	768	896	1024	1280	1536
4CIF 30Hz	1024	1280	***1536***	1792	2048
4CIF 60Hz	1536	1780	2048	2560	***3072***

In comparison to that, the layered coding case is shown as it was presented in [1]. In this case the scalable video representation is provided for the spatial and temporal resolutions and bit rates that are bold/cursive in Tab. 1. For both sequences the layered coding version typically shows slightly better results than the combined scalability version. This is mainly caused by the choice of the minimum decodable bit rates for the various spatio-temporal resolutions.

For further information, we have provided results obtained by H.264/AVC High Profile (green squares). These results are also provided for the spatial and temporal resolutions and bit rates that are bold/cursive in Tab. 1. However, in contrast to the combined scalability and layered coding cases, the classical IBBPBBP... coding temporal structure is used and some of the differences can be accounted to that. Additionally, the green squares case is also non-scalable. Note that the scalability extension of H.264/AVC also builds on top of High Profile and the optimization approach and the other parameters including motion search range are similar for all three case following [6]. While for the sequence Soccer the results of layered coding and H.264/AVC are similar, for the Crew sequence, H.264/AVC High Profile provides additional improvements indicating that further work is needed towards an efficient scalable representation.

5. CONCLUSION

A new approach to flexible combined scalability using fine granular scalability (FGS) for the scalable extension of H.264/AVC was presented. This approach was chosen as first Working Draft of the new JVT standardization activity on Scalable Video Coding. The improved scalability features are provided by introducing enhancement NAL units that contain a refinement signal for a subband picture in a coarse-to-fine representation and can be truncated at any arbitrary point. The simulation results show that the coding efficiency of this approach is only slightly worse than that of the layered representation, while it provides a much larger set of decodable spatio-temporal-rate points.

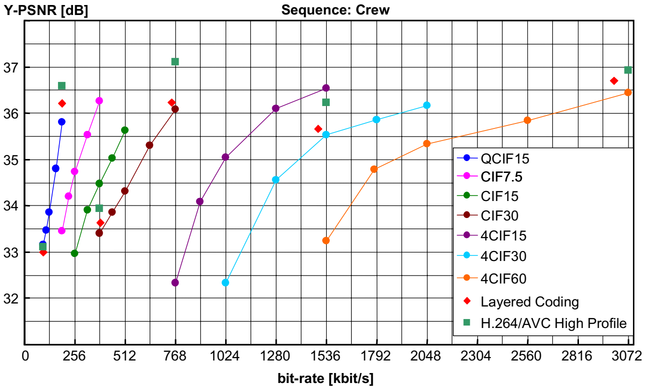

Fig. 3: PSNR vs. bit rate for the sequence Crew.

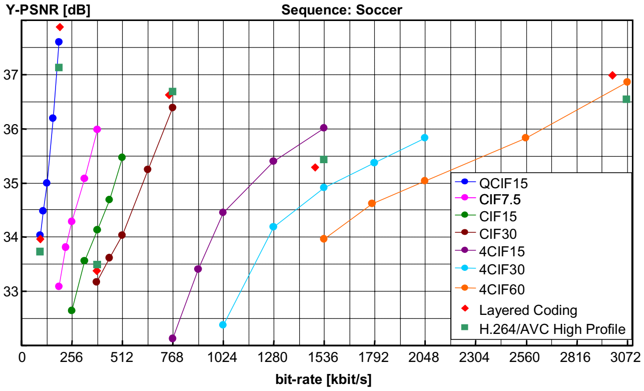

Fig. 4: PSNR vs. bit rate for the sequence Soccer.

6. REFERENCES

[1] H. Schwarz, D. Marpe, and T. Wiegand, “MCTF and Scalability Extension of H.264/AVC,” *Proc. of PCS 2004*, San Francisco, CA, USA, Dec. 2004.

[2] H. Schwarz, T. Hinz, H. Kirchhoffer, D. Marpe, and T. Wiegand, “Technical Description of the HHI proposal for SVC CE1,” ISO/IEC JTC1/SC29/WG11, Doc. m11244, Palma de Mallorca, Spain, Oct. 2004.

[3] ITU-T Recommendation H.264 & ISO/IEC 14496-10 AVC, “Advanced Video Coding for Generic Audiovisual Services,” (version 1: 2003, version 2: 2004) version 3: 2005.

[4] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. CSVT*, vol. 13, no. 7, pp. 560-576, July 2003.

[5] J. Reichel, H. Schwarz, M. Wien (eds.), “Scalable Video Coding – Working Draft 1,” *Joint Video Team (JVT)*, Doc. JVT-N020, Hong Kong, CN, Jan. 2005.

[6] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G.J. Sullivan, “Rate-Constrained Coder Control and Comparison of Video Coding Standards,” *IEEE Trans. CSVT*, vol. 13, pp. 688-703, July 2003.