SNR-SCALABLE EXTENSION OF H.264/AVC

Heiko Schwarz, Detlev Marpe, and Thomas Wiegand

Fraunhofer Institute for Telecommunications – Heinrich Hertz Institute, Image Processing Department
Einsteinufer 37, D-10587 Berlin, Germany, [hschwarz,marpe,wiegand]@hhi.fhg.de

ABSTRACT

We present an SNR-scalable extension of the H.264/AVC video coding standard. To achieve an efficient SNR-scalable bit-stream representation of a video sequence, the temporal dependencies between pictures are exploited by using an open-loop subband approach. The related temporal analysis-synthesis filter bank structure is generalized to facilitate an adaptive block-based choice between the motion-compensated lifting representations of the Haar filter (uni-directional prediction) and the 5/3 filter (bi-directional prediction), both coupled with multiple-reference frame capabilities. As a remarkable feature of our approach, most components of H.264/AVC are used as specified in the standard, while only a few have been adjusted to the motion-compensated temporal filtering structure. Our proposed SNR-scalable extension was tested for a set of CIF sequences, and the results indicate that a coding efficiency comparable to that of the state-of-the-art H.264/AVC standard can be achieved.

1. INTRODUCTION

In recent years, several efficient video codecs using motion-compensated temporal filtering have been presented [1,2,3]. The main reason for the recent advances in temporal subband coding is the utilization of the lifting representation [4] of a filter bank in the temporal direction. A two-channel decomposition can be achieved by a sequence of prediction and update steps. Since the lifting structure is invertible without requiring invertible prediction and update steps, motion-compensated prediction using any possible motion model can be incorporated into the prediction and update steps.

By using the highly efficient motion model of the H.264/AVC standard [5] in connection with an adaptive switching between the Haar and the 5/3 spline wavelet on a block basis, both the prediction and the update step are similar to the motion-compensated prediction of B slices as specified in the H.264/AVC standard. Furthermore, the open-loop structure of a temporal subband representation offers the possibility to efficiently incorporate SNR-scalability. Motivated by these facts, we have investigated the possibility of a simple but yet efficient SNR-scalable extension of H.264/AVC.

2. TEMPORAL DECOMPOSITION

In this section, we briefly review the lifting scheme and explain how it is applied to H.264/AVC video coding.

Let $s[\mathbf{x}, k]$ be a video signal with the spatial coordinate $\mathbf{x} = (x, y)^T$ and the temporal coordinate k. The decomposition of an input signal $s[\mathbf{x}, k]$ into a low-pass signal $l[\mathbf{x}, k]$ and a high-pass signal $h[\mathbf{x}, k]$ both at half the temporal resolution of the input signal is given by

$$h[\mathbf{x},k] = s[\mathbf{x},2k+1] - \mathbf{P}(s[\mathbf{x},2k+1]),$$
$$l[\mathbf{x},k] = s[\mathbf{x},2k] + \mathbf{U}(s[\mathbf{x},2k]).$$

The reconstruction of the input signal is obtained by applying the inverse operations in reverse order. The prediction and update operators for the general motion-compensated temporal filtering scheme are given by

$$\begin{aligned}\mathbf{P}_{Inter}(s[\mathbf{x},2k+1]) &= \tfrac{1}{2}(\, w_0 \cdot s[\mathbf{x}+\mathbf{m}_{P0}(\mathbf{x}),\, 2k-2r_{P0}(\mathbf{x})] + \\ &\qquad w_1 \cdot s[\mathbf{x}+\mathbf{m}_{P1}(\mathbf{x}),\, 2k+2r_{P1}(\mathbf{x})+2]\,),\\ \mathbf{U}_{Inter}(s[\mathbf{x},2k]) &= \tfrac{1}{4}(\, w_0 \cdot h[\mathbf{x}+\mathbf{m}_{U0}(\mathbf{x}),\, k+r_{U0}(\mathbf{x})] + \\ &\qquad w_1 \cdot h[\mathbf{x}+\mathbf{m}_{U1}(\mathbf{x}),\, k-r_{U1}(\mathbf{x})-1]\,),\end{aligned}$$

where $\mathbf{m}$ and r represent motion vectors and reference indices, respectively. The prediction and update operators for the motion-compensated filtering using the lifting representation of the Haar wavelet, which is given by $w_{0/1} = 2$ and $w_{1/0} = 0$, are equivalent to uni-directional motion-compensated prediction. For the 5/3 spline wavelet ($w_0 = 1$ and $w_1 = 1$), the prediction and update operators specify bi-directional motion-compensated prediction.

Since bi-directional motion-compensated prediction generally reduces the energy of the prediction residual but increases the motion information rate in comparison to uni-directional prediction, it is desirable to switch dynamically between uni- and bi-directional prediction, and thus between the lifting representation of the Haar and the 5/3 spline wavelet.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

3. INTEGRATION INTO H.264/AVC

To represent the motion fields, or more accurately the prediction data arrays $\mathbf{M}_P$ and $\mathbf{M}_U$, for the prediction and update operators, we use the existing syntax for B slices in H.264/AVC. As a slight modification, the direct macroblock and sub-macroblock mode are redefined. They specify that the corresponding macroblock or sub-macroblock is bi-directionally predicted, that the reference indices are equal to zero, and that the forward (list 0) and backward (list 1) motion vectors for the 16×16 or 8×8 block are given by the corresponding spatial motion vector predictors. Furthermore, we also incorporated an intra mode. For the intra macroblock mode, the following prediction and update operators are used

$$\mathbf{P}_{Intra}(s[\mathbf{x},2k+1])=0,$$
$$\mathbf{U}_{Intra}(s[\mathbf{x},2k])=0.$$

Thus, an intra macroblock mode in a prediction data array $\mathbf{M}_P$ specifies that in the corresponding prediction step at the analysis side, the macroblock samples of the original low-pass signal are placed into the high-pass picture. For the update step, an intra macroblock mode in a prediction data array $\mathbf{M}_U$ indicates that the update of the low-pass signal is skipped for the corresponding macroblock. Since motion vectors of the prediction data array $\mathbf{M}_U$ used in the update steps can reference an area in a high-pass picture that partially or fully covers an intra macroblock, all sample values of the intra macroblocks are set to zero for the usage in the update step.

Using the described syntax for specifying the prediction data arrays, the formation of the prediction and update pictures $\mathbf{P}(s[\mathbf{x},2k+1])$ and $\mathbf{U}(s[\mathbf{x},2k])$ is nearly identical to the motion-compensated prediction of B slices as specified in H.264/AVC.

For reducing the blocking artifacts of reconstructed pictures, the deblocking filter as specified in H.264/AVC is applied to the low-pass pictures that are reconstructed in the prediction steps at the decoder side.

In our approach, a video sequence is coded in groups of 2^N pictures, with N being the number of temporal decomposition stages. The presented two-channel decomposition is iteratively applied to the set of low-pass pictures until a single low-pass picture is obtained. The processing of a video sequence in independent groups of pictures (GOP's) generally leads to disturbing temporal blocking artifacts. In order to prevent these artifacts, we propose a temporal decomposition structure with prediction over GOP boundaries as illustrated in Fig. 1. In the prediction steps, the low-pass picture of the previous GOP that is obtained after performing all N decomposition stages is used as additional reference picture for motion-compensated prediction. However, the motion-compensated update is only performed inside the GOP; i.e. the low-pass picture of the previous GOP that is used for prediction is not updated. Note, that this decomposition structure is conceptually similar to the open-GOP structure used in hybrid video coding schemes. The temporal blocking artifacts could also be prevented by performing the temporal decomposition using a sliding window approach (cp. [1]). However, the structural delay that is associated with the sliding window approach is $3{\cdot}(2^N{-}1)$ pictures, while the structural delay of the proposed approach is only $2^N{-}1$ pictures.

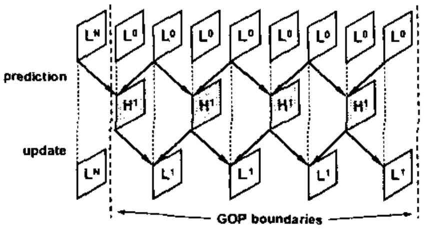

Fig. 1. Temporal decomposition with prediction over GOP boundaries.

The temporal decomposition of each group of 2^N pictures is specified by $2^N{-}1$ prediction data arrays $\mathbf{M}_P$ used in the prediction steps and $2^N{-}1$ prediction data arrays $\mathbf{M}_U$ used in the update steps. In order to reduce the bit-rate needed for transmitting the prediction data arrays, the prediction data arrays $\mathbf{M}_U$ used in the update steps are neither estimated nor coded. Instead, they are derived from the set of prediction data arrays $\mathbf{M}_P$ of the same decomposition stage. The process for deriving the prediction data arrays $\mathbf{M}_U$ is designed in a way that the derived prediction data arrays $\mathbf{M}_U$ still represent block-wise motion compatible with the B slice syntax of H.264/AVC.

In principle, the derivation process works as follows. Initially, the prediction data arrays $\mathbf{M}_U$ are divided into 4×4 blocks and for each block, a variable $N_{Covered}$ is set to zero. Then, for each motion vector $\mathbf{m}$ of the prediction data arrays $\mathbf{M}_P$, the 4×4 blocks that are at least partially covered by the area used for motion-compensated prediction of the corresponding partition are identified. If the motion vector that is already assigned to such a covered block is equal to $-\mathbf{m}$, the value of the associated variable N_{Covered} is increased by the number of newly covered samples. Otherwise, if the number of covered samples is greater than N_{Covered} for an identified block, the motion vector of the corresponding block is set equal to $-\mathbf{m}$. After processing all motion vectors of the prediction data arrays $\mathbf{M}_P$, the macroblock modes and, if appropriate, the sub-macroblocks modes of the prediction data arrays $\mathbf{M}_U$ are determined based on the motion vectors and the variables N_{Covered} of the corresponding 4×4 blocks. Due to the limitations imposed by the H.264/AVC syntax the actual process for the derivation of the prediction data arrays $\mathbf{M}_U$ is a bit more complicated, but it still follows the described principle. A detailed description of the derivation process can be found in [6].

4. SNR-SCALABLE CODING SCHEME

4.1. Base Layer Coding

For a base layer representation of a group of 2^N pictures, the 2^N-1 prediction data arrays $\mathbf{M}_P$ as well as approximations of the low-pass picture and the 2^N-1 high-pass pictures need to be transmitted. To map these data to NAL units, we use subsets of the slice layer syntax of H.264/AVC.

The prediction data arrays $\mathbf{M}_P$ are coded similar to B slices in H.264/AVC with the difference that the syntax element indicating if a macroblock is coded in skip mode is not transmitted and that no residual information (coded block pattern and residual blocks) are coded for motion-compensated macroblocks. Furthermore, only one intra mode is included in the set of possible macroblock modes. For signaling this intra mode, the codeword/binarization of the INTRA_4x4 mode is used; no intra prediction modes are transmitted. The motion vector predictors are derived as specified in the standard.

In general, a high-pass picture contains intra and residual macroblocks, where the location of intra macroblocks is specified by the corresponding prediction data array $\mathbf{M}_P$. Since the residual macroblocks represent prediction errors, the residual coding as specified in the H.264/AVC standard including transformation, scaling, and quantization is employed. For the coding of intra macroblocks, the intra macroblock modes defined in H.264/AVC can be used. However, since intra macroblocks should not be predicted from neighbouring residual macroblocks, the intra prediction is always performed as if the syntax element constrained_intra_pred_flag defined in the picture parameter set is equal to 1.

For coding the low-pass pictures, we generally use the syntax of H.264/AVC. In the simplest version, the low-pass pictures of each group of pictures (GOP) are coded independently as intra pictures. However, especially for sequences with high spatial detail and slow motion, the coding efficiency can be improved if the correlations between successive GOP's are exploited. Thus, in a more general version, low-pass pictures are coded as P pictures using reconstructed low-pass pictures of previous GOP's as references; intra (IDR) pictures are inserted in regular intervals only to provide random access points. At the decoder side, low-pass pictures are parsed and reconstructed as specified in the H.264/AVC standard including the de-blocking filter operation.

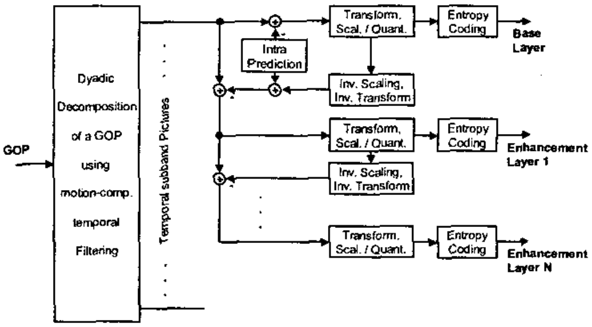
GOP
Dyadic Decomposition of a GOP using motion-comp. temporal Filtering
Temporal subband Pictures
Intra Prediction
Transform, Scal. / Quant.
Entropy Coding
Base Layer
Inv. Scaling, Inv. Transform
Transform, Scal. / Quant.
Entropy Coding
Enhancement Layer 1
Inv. Scaling, Inv. Transform
Transform, Scal. / Quant.
Entropy Coding
Enhancement Layer N

Fig. 2. Block diagram of the SNR-scalable coding scheme.

4.2. SNR-Scalability: Enhancement Layer Coding

The open-loop structure of the subband approach provides the possibility to efficiently incorporate SNR-scalability. We propose the simple but yet efficient coding scheme depicted in Fig. 2, in which the enhancement layers represent refinement pictures for the subband signals.

At the encoder side, residual pictures computed between the original subband pictures generated by the analysis filter bank and the reconstructed subband pictures obtained after decoding the previous layers, are generated. These refinement pictures are coded using the residual picture syntax. At the decoder side, the subband representation of the base layer and the refinement signals of the enhancement layers can be decoded independently. The final subband representation is obtained by adding the refinement pictures of various enhancement layers to the base layer representation.

5. EXPERIMENTAL RESULTS

For evaluating the coding efficiency of the proposed SNR-scalable extension of H.264/AVC, we compared it to an H.264/AVC compliant encoder using a similar degree of encoder optimizations. In Fig. 3, diagrams with the rate-distortion curves for the sequences "Mobile & Calendar", "Foreman", "Tempete", and "Football" are depicted.

Both encoders are operated using the Lagrangian coder control described in [7]. For the H.264/AVC compliant encoder, only the first picture is coded as IDR picture, all following pictures are coded as P and B pictures, where 2 B pictures are inserted between each pair of anchor pictures. Five reference pictures are used, and the rate-distortion curves have been obtained by varying the quantization parameter. For the scalable encoder, the GOP size was set to 32 pictures and up to 5 reference pictures have been used. The low-pass pictures of the base layer are generally coded as P pictures with the exception of the low-pass picture of the first GOP. The solid rate-distortion curves have been obtained from a single

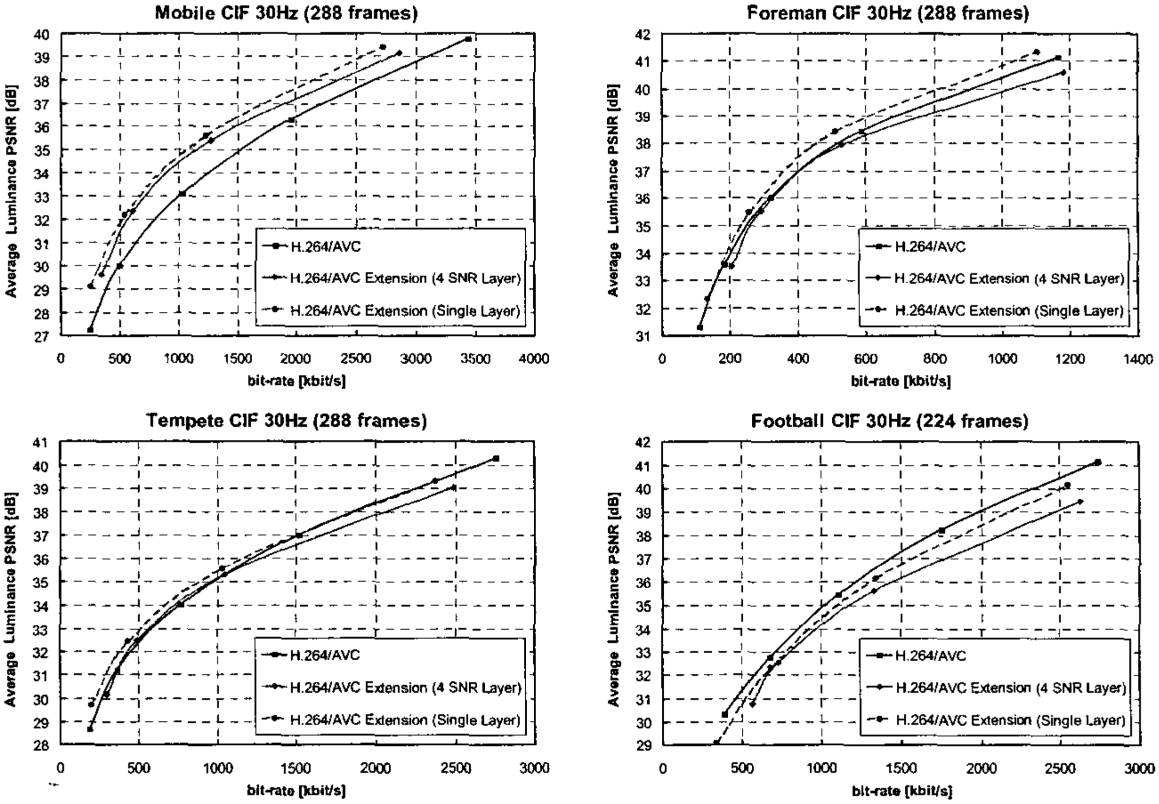

Fig. 3. Comparison of the coding efficiency of the H.264/AVC compliant coder and the proposed SNR-scalable extension for the sequences "Mobile & Calendar", "Foreman", "Tempete", and "Football" in CIF resolution (30Hz).

embedded bit-stream with 4 SNR-Layers. Additionally, the diagrams show rate-distortion curves (dashed) for a non-scalable version of the presented approach, in which only a base layer is coded; these curves have been obtained by varying the quantization parameter. For both encoders, CABAC was used as entropy coding method.

6. CONCLUSION

An extension of the H.264/AVC standard was presented that requires only a few adjustments for enabling SNR-scalability within a block-based motion-compensated temporal lifting framework. This proposed open-loop approach of motion compensation includes multiple reference frames as well as the adaptive choice between two lifting representations according to uni- and bi-directional prediction on a block basis. As a distinctive feature, motion parameters for the update process are derived from the motion parameters estimated for the corresponding prediction step. Experimental results indicate that the coding efficiency of the proposed SNR-scalable extension is comparable to that of an original H.264/AVC compliant encoder.

REFERENCES

[1] J.-R. Ohm, "Complexity and delay analysis of MCTF interframe wavelet structures," ISO/IEC JTC1/WG11 Doc. M8520, July 2002.

[2] D. Taubman, "Successive refinement of video: fundamental issues, past efforts and new directions"; *Proc. of SPIE (VCIP 2003)*, vol. 5120, pp. 649-663, July 2003.

[3] M. Flierl, "Video Coding with Lifted Wavelet Transforms and Frame-Adaptive Motion Compensation," *Proc. of VLBV*, pp. 243-251, Sep. 2003.

[4] W. Sweldens, "A custom-design construction of biorthogonal wavelets," *J. Appl. Comp. Harm. Anal.*, vol. 3 (no. 2), pp. 186-200, 1996.

[5] ITU-T Recommendation H.264 & ISO/IEC 14496-10 AVC, "Advanced Video Coding for Generic Audiovisual Services", 2003.

[6] H. Schwarz, D. Marpe, and T. Wiegand, "Scalable Extension of H.264/AVC", ISO/IEC JTC1/WG11 Doc. M10569/S03, Mar. 2004.

[7] T. Wiegand et al, "Rate-Constrained Coder Control and Comparison of Video Coding Standards," IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, pp. 688-703, July 2003.