Constrained Inter-Layer Prediction for Single-Loop Decoding in Spatial Scalability

Heiko Schwarz, Tobias Hinz, Detlev Marpe, and Thomas Wiegand
Image Processing Department
Fraunhofer Institute for Telecommunications – Heinrich Hertz Institute
Berlin, Germany
[hschwarz,hinz,marpe,wiegand]@hhi.fraunhofer.de

Abstract—**The scalability extension of H.264/AVC uses an oversampled pyramid representation for spatial scalability, where for each spatial resolution a separate motion compensation or MCTF loop is deployed. When the reconstructed signal at a lower resolution is used to predict the next higher resolution, the motion compensation or MCTF loops including the deblocking filter operations of both resolutions have to be executed. This imposes a large complexity burden on the decoding of the higher resolution signals, especially when multiple spatial layers are utilized. In this paper, we investigate the approach to only allow prediction between spatial layers for parts of the lower resolution pictures that are intra-coded in order to avoid decoding that requires multiple motion compensation or MCTF loops. Experimental results evaluate the effectiveness of the proposed approach.**

Keywords–**Scalability; H.264/AVC; inter-layer prediction; single-loop decoding**

I. 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, the ISO/IEC Moving Pictures Experts Group (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]. A reference encoder is described in the Joint Scalable Video Model 0 (JSVM 0) [6].

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. The basic design of the scalable extension of H.264/AVC can be classified as layered video codec. In each layer, the basic concepts of motion-compensated prediction and intra prediction are employed as in standard H.264/AVC. However, additional inter-layer prediction mechanisms have been integrated in order to exploit the redundancy between several layers. SNR scalability is basically achieved by residual quantization with little changes to H.264/AVC. For spatial scalability, a combination of motion-compensated prediction and oversampled pyramid decomposition is proposed, which requires some additional mechanisms to convey bit-rate from the lower resolution to the higher resolution layers. Because of the similarities in motion-compensated prediction, the approach to temporal scalability of H.264/AVC is maintained.

Among the various types of scalability, spatial scalability requires the largest degree of change to H.264/AVC. Various mechanisms to re-use information from a lower spatial resolution in a higher spatial resolution layer are specified. One of these mechanisms is the upsampling of the decoded signal at the lower resolution and making this signal available for prediction. However, this feature has the drawback that both motion compensation loops of the lower and higher resolution must be executed for parts of the picture that are not intra-coded in the lower resolution signal. In this paper, we investigate an approach that enables a single-loop decoding by constraining the inter-layer prediction.

The next section outlines the basic concepts of the scalability extension of H.264/AVC while a more detailed description can be found in [1][2][5]. Section III describes the constrained inter-layer prediction for enabling single-loop decoding, and Section IV provides experimental results comparing single-loop and multiple-loop decoding.

II. SCALABLE EXTENSION OF H.264/AVC

The scalable H.264/AVC extension specifies a layered video codec. In general, the coder structure depends on the scalability space that is required by the application. For illustration, Fig. 1 shows a typical coder structure with 2 spatial layers. In each layer, which can either be a spatial layer or a coarse-grain SNR layer, an independent hierarchical motion-compensated prediction structure with layer-specific motion parameters is employed. The redundancy between consecutive layers is exploited by different inter-layer prediction concepts that include prediction mechanisms for motion parameters as well as texture data. A base representation of the input pictures of each layer is obtained by transform coding similar to that of H.264/AVC, the corresponding NAL units (NAL – Network Abstraction Layer) contain motion information and texture data; the NAL units of the base representation of the lowest layer are compatible with standard H.264/AVC. The reconstruction quality of the base representations can be improved by an additional coding of so-called progressive refinement slices; the corresponding NAL can be arbitrarily truncated in order to support fine granular quality scalability (FGS) or flexible bit-rate adaptation.

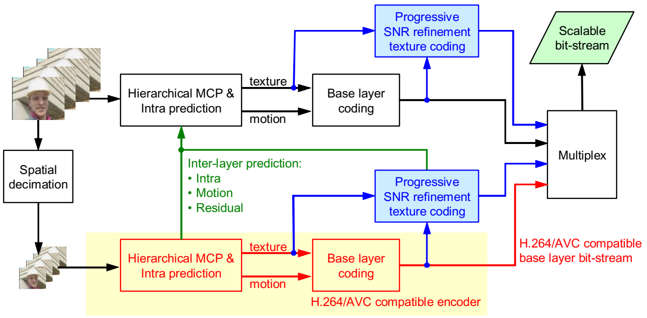
Progressive SNR refinement texture coding
Scalable bit-stream
texture
Hierarchical MCP & Intra prediction
Base layer coding
motion
Multiplex
Spatial decimation
Inter-layer prediction:
• Intra
• Motion
• Residual
Progressive SNR refinement texture coding
texture
Hierarchical MCP & Intra prediction
Base layer coding
motion
H.264/AVC compatible base layer bit-stream
H.264/AVC compatible encoder

Fig. 1. Typical coder structure for the scalable extension of H.264/AVC.

A. *Hierarchical Prediction Structure and Temporal Scalability*

Generally, in each layer a hierarchical prediction structure as illustrated in Fig. 2 is employed. The first picture of a video sequence is coded as IDR picture; so-called key pictures are coded in regular intervals. A key picture and all pictures that are temporally located between a key picture and the previous key pictures are considered to build a group of pictures (GOP). The key pictures are either intra-coded or inter-coded by using previous key pictures as reference for motion-compensated prediction. The remaining pictures of a GOP are hierarchically predicted as shown in Fig. 2. It is obvious that this hierarchical prediction structure inherently provides temporal scalability; but it turned out that it also offers the possibility to efficiently integrating SNR scalability.

The hierarchical picture coding can be extended to motion-compensated filtering (MCTF). For that, motion-compensated update operations using the prediction residuals (dashed arrows in Fig. 2) are introduced in addition to the motion-compensated prediction. For more details on how the hybrid coding approach of H.264/AVC is extended towards MCTF please refer to [1][2][5][6].

B. *SNR Scalability*

For the SNR base layer (base representation), H.264/AVC-conforming transform coding is used. For each macroblock, the coded block pattern (CBP), and conditioned on CBP the corresponding residual blocks are transmitted together with the macroblocks modes, intra prediction modes, and motion data using the I, P, or B slice syntax of H.264/AVC.

On top of the SNR base layer, SNR enhancement layers are coded. For that, the quantization error between the SNR base layer and the original residual and intra macroblocks is re-quantized exactly using the same methods as for the base layer but with a finer quantization step size, i.e., a lower value of the quantization parameter. In a simple version, the transform coefficient levels of the SNR enhancement layers are transmitted using the residual syntax of H.264/AVC. With this approach only coarse grains of scalable SNR layers can be efficiently represented as factors of 2 in bit-rate. In order to support fine granular SNR scalability, we have introduced so-called progressive refinement slices, in which the symbol coding order has been changed in a way that the corresponding NAL units can be truncated at any arbitrary point.

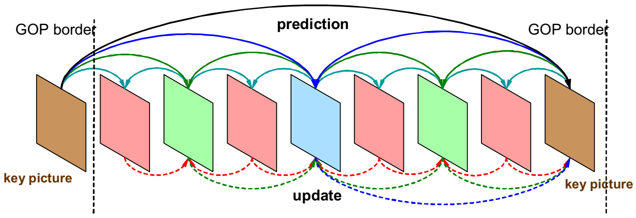

Fig. 2. Hierarchical prediction structure.

Note, that it is basically also possible to specify motion field refinements for SNR enhancement layers. Therefore, the same inter-layer prediction techniques as described in the next subsection but without the upsampling operations are applied.

C. *Inter-Layer Prediction and Spatial Scalability*

As a first interpretation, the pictures (base representations) for different layers are coded independently with layer-specific motion information. We consider spatial scalability with a factor of 2 in horizontal and vertical resolution, although the concepts can be generalized. From several experiments we have found that it would be efficient to allow the encoder to freely choose which dependencies between spatial resolution layers need to be exploited through switchable prediction mechanisms. The following techniques turned out to provide gains and were included into the scalable video codec:

- Prediction of motion vectors using the upsampled lower resolution motion vectors
- Prediction of the residual signal using the upsampled residual signal of the lower resolution layer
- *Prediction of a macroblock using the reconstructed and upsampled lower resolution signal*

The last of the 3 methods is the one modified in this work while the other two remain unchanged. The inter-layer prediction techniques are briefly described in the following.

1) Motion Vector Prediction: For prediction of motion vectors using the upsampled lower resolution motion vectors we have introduced two additional macroblock modes that utilize motion information of the lower resolution layer. The macroblock partitioning is obtained by upsampling the partitioning of the corresponding 8x8 block of the lower resolution layer. For the obtained macroblock partitions, the same reference picture indices as for the corresponding sub-macroblock partition of the base layer block are used; and the associated motion vectors are scaled by a factor of 2. While for the first of these macroblock modes no additional motion information is coded, for the second one, a quarter-sample motion vector refinement is transmitted for each motion vector. Additionally, our approach includes the possibility to use a scaled motion vector of the lower resolution as motion vector predictor for the conventional motion-compensated macroblock modes.

2) Residual Prediction: In order to also incorporate the possibility of exploiting the residual information coded in the lower resolution layer, an additional flag is transmitted for each macroblock, which signals the application of residual signal prediction from the lower resolution layer. If the flag is true, the base layer residual signals is block-wise up-sampled using a bi-linear filter with constant border extension and used as prediction for the residual signal of the current layer, so that only the corresponding difference signal is coded.

3) Intra Prediction: We have further introduced an additional intra macroblock mode. In that mode, the intra prediction signal is generated by upsampling the reconstruction signal of the lower resolution layer using the 6-tap filter which is specified in H.264/AVC for the purpose of half-sample interpolation. The prediction residual is transmitted using H.264/AVC residual coding.

III. Constrained Inter-Layer Prediction

For the inter-layer prediction using the reconstructed lower resolution signal as described in Sec. II.C-3 it is generally required that the lower resolution layer is completely decoded including the computationally complex operations of motion-compensated prediction (or inverse MCTF) and deblocking. Fig. 3 illustrates the problem. The yellow-marked pictures are layer *k* pictures of the pyramid and need to be decoded using motion-compensated prediction (or inverse MCTF) and deblocking. The blue-marked pictures represent the upsampled versions of the decoded layer *k* pictures. For decoding layer *k+1*, the orange-marked macroblocks are predicted from the decoded and upsampled pictures of layer *k*. It is worth noting that at the decoder only those parts of layer *k* need to be upsampled that are actually used for prediction. However, the main problem remains that in general the motion-compensated prediction (or inverse MCTF) as well as the deblocking for layer *k* must be executed to decode layer *k+1*. This creates a large complexity overhead for the decoding process of layer *k+1* and even more for all higher layers.

We have found that the above problem can be circumvented by restricting the prediction from upsampled decoded pictures to those parts of the lower layer picture that are coded with intra macroblocks. For that, the intra prediction signal is directly obtained by deblocking and upsampling the corresponding 8x8 luma block inside the corresponding lower layer picture. With the proposed changes, the decoding complexity is significantly reduced, since the motion-compensated prediction (or inverse MCTF) as well as the deblocking of inter-coded macroblocks is only required for the spatial or coarse-grain SNR layer that is actually decoded.

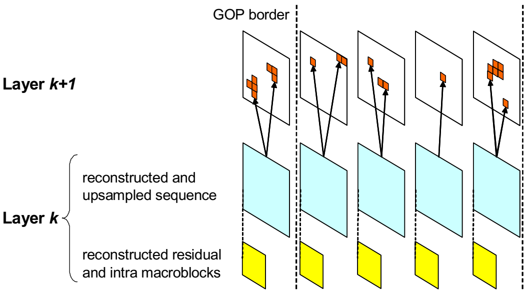
GOP border
Layer *k+1*
Layer *k*
reconstructed and upsampled sequence
reconstructed residual and intra macroblocks

Fig. 3. Unrestricted inter-layer prediction of intra macroblocks.

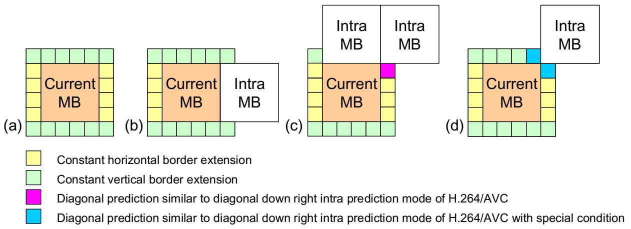
Current MB
Intra MB
(a)
(b)
(c)
(d)
Constant horizontal border extension
Constant vertical border extension
Diagonal prediction similar to diagonal down right intra prediction mode of H.264/AVC
Diagonal prediction similar to diagonal down right intra prediction mode of H.264/AVC with special condition

Fig. 4. Padding of intra macroblocks before upsampling.

The interpolation of an 8x8 block of the lower layer is generally performed using the half-pel interpolation filter of H.264/AVC. Before interpolation, the block edges inside intra macroblocks as well as the macroblock edges between intra macroblocks of the base layer are deblocked as specified in H.264/AVC, and afterwards these modified intra macroblocks are extended by a 4-pixel border in each direction using the following padding process (see Fig. 4).

When a neighboring macroblock is coded in an intra mode no border extension is performed but the corresponding intra samples are used for interpolation (Fig. 4b,c,d). Otherwise, the 4-pixel border is generally obtained by horizontal constant border extension of the current macroblock (as well as of vertical neighbors coded in intra mode) and a subsequent vertical border extension. However, if a horizontal or vertical neighboring macroblock is coded in an inter mode but one of the two diagonal neighboring macroblocks is coded in intra mode, then the 4x4 block of the border extension that is located next to the diagonal intra-coded neighbor (e.g. magenta colored block in Fig. 4c) is generated similar to the intra prediction signal for the diagonal down right intra prediction mode of H.264/AVC. When the corresponding 4x4 block does not represent the upper block of the right border, the coordinates that are used in the intra prediction process are modified accordingly. If the corner sample that is needed for generating the intra prediction signal does not belong to an intra coded macroblock as in Fig. 4d, it is replaced by the average of the two neighboring corner samples.

IV. Experimental Results

For evaluating the impact of the proposed constrained inter-layer prediction on coding efficiency, we compared it with the general scheme [1] that requires a multiple-loop decoding. Note that multiple-loop decoding stands for the unrestricted version of prediction of a macroblock using the reconstructed and upsampled lower resolution signal. We have chosen a scalability scenario that consists of 5 spatial or coarse-grain SNR layers. The lowest layer is coded conforming with the H.264/AVC standard; for all enhancement layers MCTF has been applied. With the general multiple-loop scheme, 5 motion compensation or MCTF loops are required for decoding the highest layer; whereas with the proposed modification, the decoding of the highest layer can be realized with a single MCTF loop. For both codec versions the same encoder control following [8] was used.

Our simulation results have shown that for most sequences, the impact on coding efficiency is small by imposing our proposed restriction. An example for such a case which stands

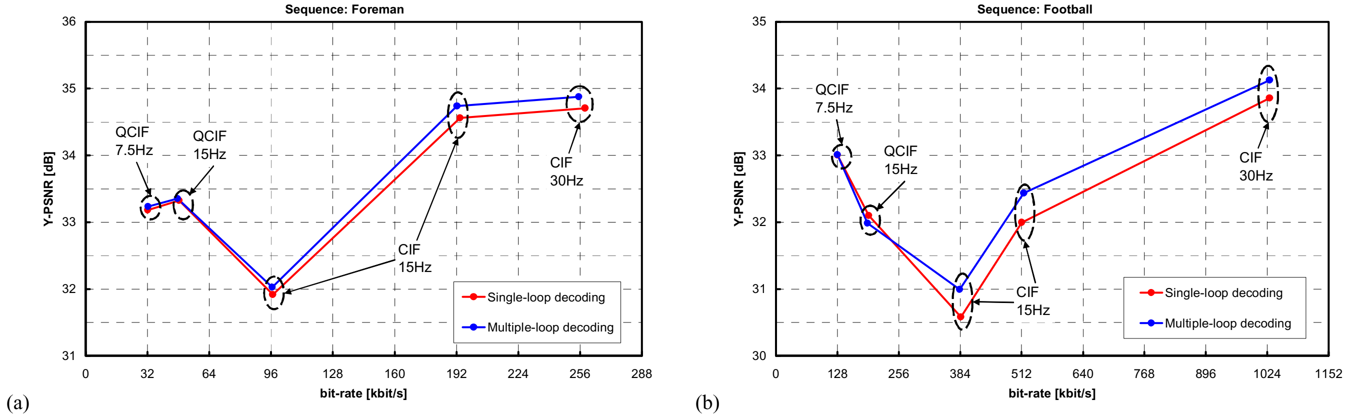

Fig. 5. Coding results for the sequences "Foreman" and "Football".

for a larger number of other cases is shown in Fig. 5a for the Foreman sequence. In addition to the many cases for which we have found small impairments due to our proposed restriction leading to single-loop decoding, we have found two sequences, namely Football and Crew, for which we have observed a more significant influence on the coding efficiency. As an example, the result for Football is depicted in Fig. 5b.

In Fig. 6, the usage of macroblock modes for the first CIF 15Hz layer has been analyzed for the sequences Foreman and Football. "Intra" represents the intra modes of standard H.264/AVC, while "Intra_BL" stands for the intra mode that uses the upsampled base layer signal as prediction. "BLMode" stands for the macroblock modes that employ the motion partitioning, the reference indices, and motion vectors of the base layer (cp. Sec. II.C-1), while the motion-compensated macroblock modes of standard H.264/AVC are labeled by the term "NormalMC". The suffix "+RP" indicates the additional usage of inter-layer residual prediction (cp. Sec. II.C-2) for the motion-compensated macroblock modes.

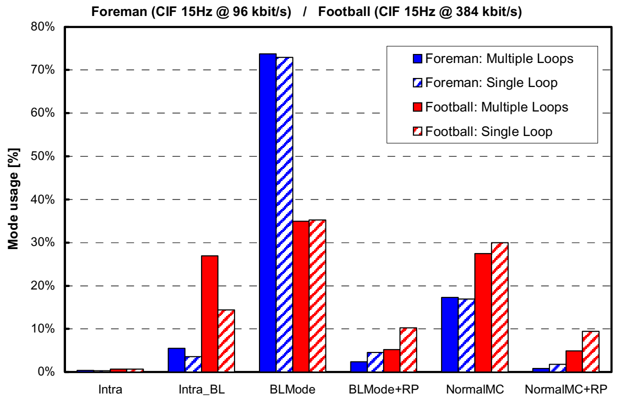

Fig. 6. Mode usage for the first CIF 15Hz layer.

The analysis shows that the largest impairments of the coding efficiency are observed for sequences, for which a large degree of macroblocks (about 27% for the Football sequence) are coded using the "Intra_BL" mode with the general unrestricted inter-layer prediction scheme. By constraining the usage of the "Intra_BL" mode, a major part of these macroblocks is inter-coded, mainly with additional residual prediction from the base layer. The sequences, for which a significant impairment of the coding efficiency is observed, are mainly characterized by fast and complex motion that cannot be well presented by motion-compensated prediction.

V. Conclusion

We have presented a simple modification to the inter-layer prediction in pyramid-based spatial scalability. The approach is integrated into the scalability extension of H.264/AVC which was chosen as the first Working Draft of the new JVT standardization activity on Scalable Video Coding. By restricting the prediction to upsampled intra-coded image parts, a single-loop decoder for spatial and coarse-grain SNR scalability can be realized requiring only the motion-compensated prediction or inverse MCTF as well as deblocking of inter-coded macroblocks for the scalable layer that is being decoded. The rate-distortion penalty for this restriction is found for most sequences to be small while only a few sequences are found with PSNR losses up to 0.7 dB.

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 CE 1," ISO/IEC JTC1/WG11, Doc. m11244, Palma de Mallorca, Spain, Oct. 2004.

[3] ITU-T Recommandation 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. CVST, vol. 13, no. 7, pp. 560-576, July 2003.

[5] J. Reichel, H. Schwarz, and M. Wien (eds.), "Scalable Video Coding – Working Draft 1," Joint Video Team of ITU-T VCEG and ISO/IEC MPEG, Doc. JVT-N020, Hong Kong, CN, Jan. 2005.

[6] J. Reichel, H. Schwarz, and M. Wien (eds.),"Joint Scalable Video Model (JSVM) 0," Joint Video Team of ITU-T VCEG and ISO/IEC MPEG, Doc. JVT-N021, Hong Kong, CN, Jan. 2005.

[7] H. Schwarz, D. Marpe, and T. Wiegand, "Further results on constrained inter-layer prediction," Joint Video Team, Doc. JVT-O074, Busan, KR, April 2005.

[8] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-Constrained Coder Control and Comparison of Video Coding Standards," IEEE Trans. CVST, vol. 13, no. 7, pp. 688-703, July 2003.