OVERVIEW OF THE SCALABLE H.264/MPEG4-AVC EXTENSION

Heiko Schwarz, Detlev Marpe, and Thomas Wiegand

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

ABSTRACT

The scalable extension of H.264/MPEG4-AVC is a current standardization project of the Joint Video Team of the ITU-T Video Coding Experts Group and the ISO/IEC Moving Picture Experts Group. This paper gives an overview of the design of the scalable H.264/MPEG4-AVC extension and describes the basic concepts for supporting temporal, spatial, and SNR scalability. The efficiency of the described concepts for providing spatial and SNR scalability is analyzed by means of simulation results and compared to H.264/MPEG4-AVC compliant single layer coding.

Index Terms— video coding

1. INTRODUCTION

Scalable Video Coding (SVC) is currently a very active working area in the research community and in international standardization. A project on SVC standardization was originally started by the ISO/IEC Moving Picture Experts Group (MPEG). Based on an evaluation of the submitted proposal, MPEG and the ITU-T Video Coding Experts Group (VCEG) agreed to jointly finalize the SVC project as an Amendment of their H.264/MPEG4-AVC standard [1], for which the scalable extension of H.264/MPEG4-AVC as proposed in [2] was selected as the first Working Draft.

As an important feature of the SVC design, most components of H.264/MPEG4-AVC are used as specified in the standard. This includes the motion-compensated and intra prediction, the transform and entropy coding, the deblocking as well as the NAL unit packetization (NAL – Network Abstraction Layer). The base layer of an SVC bit-stream is generally coded in compliance with H.264/MPEG4-AVC, and each standard conforming H.264/MPEG-4 AVC decoder is capable of decoding this base layer representation when it is provided with an SVC bit-stream. New tools are only added for supporting spatial and SNR scalability.

This paper gives an overview of the current SVC design [3][4]. The basic concepts for proving temporal, spatial, and SNR scalability are described and analyzed regarding their coding efficiency. For more detailed information, the reader is referred to the SVC Working Draft [3] and the Joint Scalable Video Model (JSVM) [4].

2. OVERVIEW

The basic SVC design can be classified as layered video codec. In general, the coder structure as well as the coding efficiency depends on the scalability space that is required by an application. For illustration, Fig. 1 shows a typical coder structure with two spatial layers.

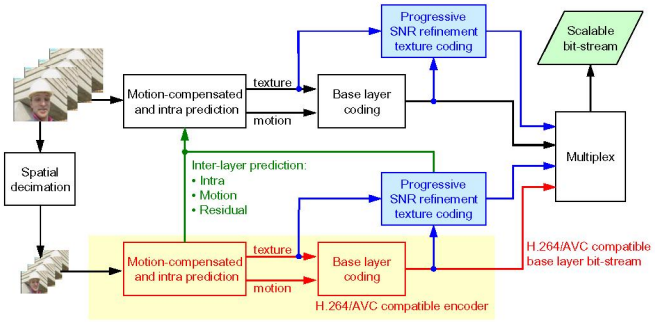

Fig. 1. Coder structure example with two spatial layers.

In each spatial or coarse-grain SNR layer, the basic concepts of motion-compensated prediction and intra prediction are employed as in H.264/MPEG4-AVC. The redundancy between different layers is exploited by additional inter-layer prediction concepts that include prediction mechanisms for motion parameters as well as texture data (intra and residual data). A base representation of the input pictures of each layer is obtained by transform coding similar to that of H.264/MPEG4-AVC, the corresponding NAL units contain motion information and texture data; the NAL units of the lowest layer are compatible with single-layer H.264/MPEG4-AVC [1]. The reconstruction quality of these base representations can be improved by an additional coding of so-called progressive refinement slices. In contrast to all other slice data NAL units, the corresponding NAL units can be arbitrarily truncated in order to support fine granular quality scalability or flexible bit-rate adaptation.

An important feature of the SVC design is that scalability is provided at a bit-stream level. Bit-streams for a reduced spatial and/or temporal resolution are simply obtained by discarding NAL units (or network packets) from a global SVC bit-stream that are not required for decoding the target resolution. NAL units of PR slices can additionally be truncated in order to further reduce the bit-rate and the associated reconstruction quality.

3. TEMPORAL SCALABILITY AND HIERARCHICAL CODING STRUCTURES

In contrast to older video coding standards as MPEG-2/4, the coding and display order of pictures is completely decoupled in H.264/MPEG4-AVC. Any picture can be marked as reference picture and used for motion-compensated prediction of following pictures independent of the corresponding slice coding types. These features allow the coding of picture sequences with arbitrary temporal dependencies.

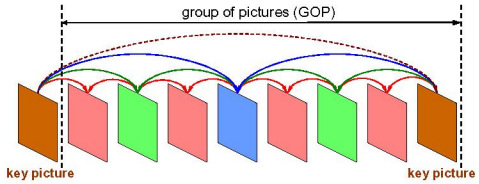

Fig. 2. Hierarchical prediction structure with 4 dyadic levels.

Temporal scalable bit-stream can be generated by using hierarchical prediction structures as illustrated in Fig. 2 without any changes to H.264/MPEG4-AVC. So-called key pictures are coded in regular intervals by using only previous key pictures as references. The pictures between two key pictures are hierarchically predicted as shown in Fig. 2. It is obvious that the sequence of key pictures represents the coarsest supported temporal resolution, which can be refined by adding pictures of following temporal prediction levels.

In addition to enabling temporal scalability, the hierarchical prediction structures also provide an improved coding efficiency compared to classical IBBP coding on the cost of an increased encoding-decoding delay [5]. Furthermore, the efficiency of the tools for supporting spatial and SNR scalability is improved as it will be proven in the following sections. It should also be noted that the delay of hierarchical prediction structures can be controlled by restricting the motion-compensated prediction from pictures of the future.

4. SPATIAL SCALABILITY

Spatial scalability is achieved by an oversampled pyramid approach. The pictures of different spatial layers are independently coded with layer specific motion parameters as illustrated in Fig. 1. However, in order to improve the coding efficiency of the enhancement layers in comparison to simulcast, additional inter-layer prediction mechanisms have been introduced. These prediction mechanisms have been made switchable, so that an encoder can freely choose which base layer information should be exploited for an efficient enhancement layer coding. Since the incorporated inter-layer prediction concepts include techniques for motion parameter and residual prediction, the temporal prediction structures of the spatial layers should be temporally aligned for an efficient use of the inter-layer prediction. It should be noted that all NAL units for a time instant form an access unit and thus have to be follow each other inside an SVC bit-stream.

4.1. Inter-layer prediction techniques

The following three inter-layer prediction techniques are included in the SVC design. In the following, only the original concepts based on simple dyadic spatial scalability are described. For an extension to arbitrary resolution ratios the reader is referred to [3][4][6].

4.1.1. Inter-layer motion prediction

In order to employ base layer motion data for spatial enhancement layer coding, additional macroblock modes have been introduced in spatial enhancement layers. The macroblock partitioning is obtained by upsampling the partitioning of the co-located 8x8 block in the lower resolution layer. The reference picture indices are copied from the co-located base layer blocks, and the associated motion vectors are scaled by a factor of 2. These scaled motion vectors are either used unmodified or refined by an additional quarter-sample motion vector refinement. Additionally, a scaled motion vector of the lower resolution can be used as motion vector predictor for the conventional macroblock modes.

4.1.2. Inter-layer residual prediction

The usage of inter-layer residual prediction is signaled by a flag that is transmitted for all inter-coded macroblocks. When this flag is true, the base layer signal of the co-located block is block-wise upsampled and used as prediction for the residual signal of the current macroblock, so that only the corresponding difference signal is coded.

4.1.3. Inter-layer intra prediction

Furthermore, an additional intra macroblock mode is introduced, in which the prediction signal is generated by upsampling the co-located reconstruction signal of the lower layer. For this prediction it is generally required that the lower layer is completely decoded including the computationally complex operations of motion-compensated prediction and deblocking. However, as shown in [7] this problem can be circumvented when the inter-layer intra prediction is restricted to those parts of the lower layer picture that are intra-coded. With this restriction, each supported target layer can be decoded with a single motion compensation loop.

4.2. Performance evaluation

The performance of the spatial scalability tools has been evaluated in comparison to simulcast and single-layer coding. The base layer was coded at a fixed bit-rate, for the encoding of the spatial enhancement layers, the bit-rate as well as the amount of enabled inter-layer prediction mechanisms was varied. All encoders have been rate-distortion optimized following [8]. The intra period was set to 32 pictures; simulations have been carried out for a GOP size of 16 pictures as well as for IPPP coding. In Fig. 3, the results for the sequence "Soccer" with a CIF and a 4CIF layer are shown.

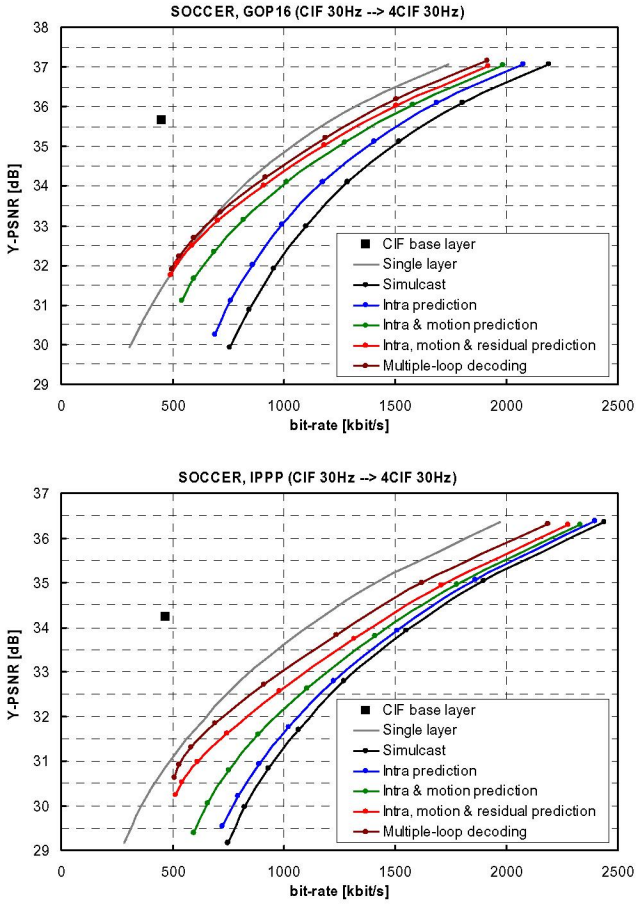

Fig. 3. Performance of inter-layer prediction mechanisms.

The black and the grey curve represent single-layer coding and simulcast, respectively. For the blue, green, and red curve, the inter-layer intra, motion, and residual prediction have been successively enabled. For all these curves, the inter-layer prediction was restricted in a way that allows single-loop decoding. The efficiency of spatial scalable coding that requires multiple-loop decoding is represented by the brown curve. By comparing both diagrams of Fig. 3 it can be seen that the efficiency of the inter-layer prediction is improved by using hierarchical prediction structures.

5. SNR SCALABILITY

For SNR scalability, coarse-grain scalability (CGS) and fine-grain scalability (FGS) are distinguished.

5.1. Coarse-grain SNR scalability

Coarse-grain SNR scalable coding is achieved using the concepts for spatial scalability. The only difference is that for CGS the upsampling operations of the inter-layer prediction mechanisms are omitted. Note that the restricted inter-layer prediction that enables single-loop decoding is even more important for CGS than for spatial scalable coding.

5.2. Fine-grain SNR scalability

In order to support fine-granular SNR scalability, so-called progressive refinement (PR) slices have been introduced. Each PR slice represents a refinement of the residual signal that corresponds to a bisection of the quantization step size (QP increase of 6). These signals are represented in a way that only a single inverse transform has to be performed for each transform block at the decoder side. The ordering of transform coefficient levels in PR slices allows the corresponding PR NAL units to be truncated at any arbitrary byte-aligned point, so that the quality of the SNR base layer can be refined in a fine-granular way.

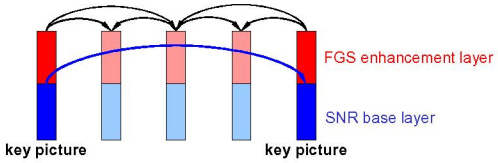

Fig. 4. Motion-compensated prediction with FGS.

The main reason for the low performance of the FGS in MPEG-4 is that the motion-compensated prediction (MCP) is always done in the SNR base layer. In the SVC design, the highest quality reference available is employed for the MCP of non-key pictures as depicted in Fig. 4. Note that this difference significantly improves the coding efficiency without increasing the complexity when hierarchical prediction structures are used. The MCP for key pictures is done by only using the base layer representation of the reference pictures. Thus, the key pictures serve as re-synchronization points, and the drift between encoder and decoder reconstruction is efficiently limited.

In order to improve the FGS coding efficiency, especially for low-delay IPPP coding, leaky prediction concepts for the motion-compensated prediction of key pictures have been additionally incorporated in the SVC design [3][4][9]. In [10], a method for further improving the FGS coding efficiency by allowing the coding of motion parameter refinements as part of the PR slices has been proposed.

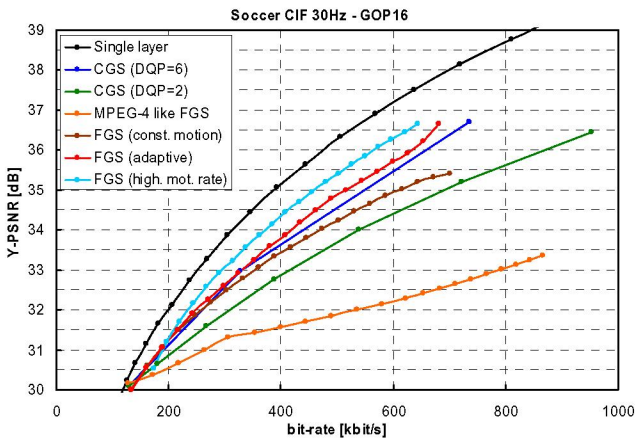

Fig. 5. Performance of SNR scalable coding strategies.

5.3. Performance evaluation

The performance of the different presented SNR scalable coding strategies have been compared to single-layer coding. Only the first picture has been intra-coded, and a GOP size of 16 pictures was chosen. The difference between the quantization parameters of the lowest and highest SNR layer was set to 12. The simulation results for "Soccer" in CIF resolution and a frame rate of 30Hz are depicted in Fig. 5.

The black curve represent the coding efficiency of single-layer coding. The blue and green curve represent CGS runs with quantization parameter differences between successive layer of 6 and 2, respectively. It is clearly visible that the coding efficiency of CGS decreases with smaller bit-rate ratios between successive SNR layers. The performance of MPEG4-like FGS coding, in which only the SNR base layer is used for MCP, is shown by the orange curve. The FGS coding efficiency can be significantly improved when higher quality references are used for the prediction of non-key pictures as specified in the SVC design and shown by the brown curve. The refinement of motion parameters in PR slices as proposed in [10] lead to further improvements of the FGS coding efficiency as illustrated by the red curve. By means of the light blue curve, it is illustrated on the example of the adaptive FGS concept [10] how the coding efficiency can be traded-off for low and high rates by modifying the ratio between motion and texture rate.

6. COMBINED SCALABILITY

The presented concepts for temporal, spatial, and SNR scalability can be easily combined. In Fig. 6, the coding efficiency of combined scalable coding is compared to the coding efficiency of single-layer, purely spatial scalable, and purely SNR scalable coding for the sequence "Soccer". The intra period was generally set 1.07 s (64 pictures at 60Hz); and for all encodings a dyadic hierarchical prediction structure with a GOP size of 32 pictures at 60Hz has been used. Since, temporal scalability with resolution of 1.875Hz to 60Hz (4CIF) is supported in the same manner in all bit-streams, it has not been tested separately.

The black curves show the coding efficiency of single-layer coding; each point represents a separate bit-stream. For SNR scalability, a separate bit-stream has been generated for each spatial resolution and is represented by the corresponding red curves. Similarly, the blue curves show the coding efficiency of spatial scalable coding. Three bit-stream have been generated, and each of these bit-streams includes either the lowest, the middle, or the highest plotted rate point for each spatial resolution. The coding efficiency of combined scalable coding, for which all plotted spatio-temporal rate points are supported in a single bit-stream is represented by the green curves. As it can be seen on the example of Fig. 6, the coding efficiency of SVC bit-streams scales with the range of supported spatio-temporal-rate points.

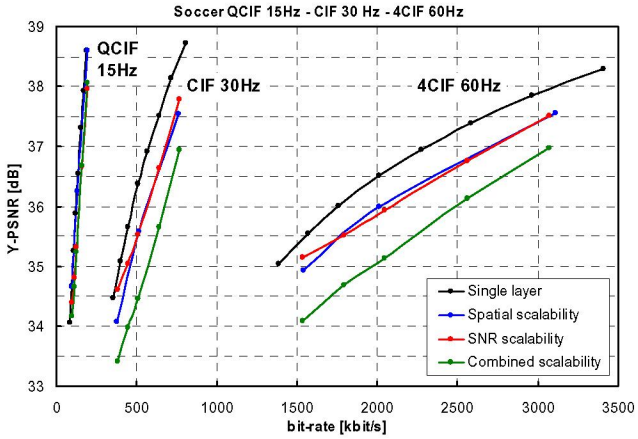

Fig. 6. Performance of combined scalability coding.

7. CONCLUSION

In this paper, the design of the scalable H.264/MPEG4-AVC extension is described, and the basic tools for proving spatial and SNR scalability are analyzed regarding their coding efficiency. The coding efficiency of SVC bit-streams and the amount of supported spatio-temporal-rate points can be traded-off according to the needs of an application.

REFERENCES

[1] ITU-T Rec. & ISO/IEC 14496-10 AVC, "Advanced Video Coding for Generic Audiovisual Services," version 3, 2005.

[2] H. Schwarz, et. al, "Technical Description of the HHI proposal for SVC CE1," *ISO/IEC JTC1/WG11*, Doc. m11244, Palma de Mallorca, Spain. Oct. 2004.

[3] J. Reichel, H. Schwarz, and M. Wien (eds.), "Scalable Video Coding – Joint Draft 4," Joint Video Team, Doc. JVT-Q201, Nice, France, Oct. 2005.

[4] J. Reichel, H. Schwarz, and M. Wien (eds.), "Joint Scalable Video Model JSVM-4," Joint Video Team, Doc. JVT-Q202, Nice, France, Oct. 2005.

[5] H. Schwarz, D. Marpe, and T. Wiegand, "Hierarchical B pictures," Joint Video Team, Doc. JVT-P014, Poznan, Poland, July 2005.

[6] E. François and J. Vieron, "Extended spatial scalability: a generalization of spatial scalability for non-dyadic configurations," *submitted to ICIP 2006.*

[7] H. Schwarz, T. Hinz, D. Marpe, and T. Wiegand, "Constrained inter-layer prediction for single-loop decoding in spatial scalability," *Proc. of ICIP 2005*, Genova, Italy, Sep. 2005.

[8] T. Wiegand, et. al, "Rate-constrained coder control and comparison of video coding standards," *IEEE Trans. CSVT*, vol. 13, pp. 688-703, July 2003.

[9] J. Ridge, X. Wang, Y. Bao, and M. Karczewicz, "Low-delay, low-complexity scalable bit-rate video coding," *submitted to ICIP 2006.*

[10] M. Winken, H. Schwarz, D. Marpe, and T. Wiegand, "Adaptive motion refinement for FGS slices," Joint Video Team, Doc. JVT-Q031, Nice, France, Oct. 2005.