Performance Analysis of SVC

Mathias Wien, *Member, IEEE*, Heiko Schwarz, and Tobias Oelbaum

(Invited Paper)

***Abstract*—This paper provides a performance analysis of the Scalable Video Coding (SVC) extension of H.264/AVC. A short overview presenting the main functionalities of SVC is given and main issues in encoder control and bit stream extraction are outlined. Some aspects of rate-distortion optimization in the context of SVC are discussed and strategies for derivation of optimized configurations relative to the investigated scalability scenarios are presented. Based on these methods, rate-distortion results for several SVC configurations are presented and compared to rate-distortion optimized H.264/AVC single layer coding. For reference, a comparison to rate-distortion optimized MPEG-4 Visual (Advanced Simple Profile) coding results is provided. The results show that the performance gap between single layer coding and scalable video coding can be very small and that SVC clearly outperforms previous video coding technology such as MPEG-4 ASP.**

***Index Terms*—H.264/AVC, rate-distortion optimization, performance, scalability, video coding.**

I. Introduction

SCALABLE profiles have been developed for multiple older video coding standards [1]–[3]. However, especially for quality scalability, the rate-distortion efficiency of these schemes was limited when compared to nonscalable single layer coding [4]. With the Scalable Video Coding (SVC) extension of H.264/AVC, the gap in rate-distortion performance between state-of-the-art single-layer coding and scalable coding could be significantly reduced. The improved coding efficiency of SVC relative to the scalable profiles of previous standards can be attributed to the possibility of using efficient hierarchical prediction structures, the new inter-layer prediction mechanisms, the improved drift control for quality scalable coding with packet-based granularity as well as the efficient coding tools of H.264/AVC such as the variable block size motion-compensated prediction with multiple reference pictures or the efficient entropy coding methods.

SVC supports scalability in terms of spatial and temporal resolution as well as the variation of the reconstruction quality. While temporal scalability can be efficiently provided by the concept of hierarchical temporal prediction structures as already supported in the nonscalable profiles of H.264/AVC, the scalability features in terms of spatial resolution and fidelity enhancements require the introduction of new tools into the existing single layer coding scheme of H.264/AVC. Spatial scalability is supported for dyadic resolution ratios as well as for generalized relationships between spatial layers. The generalized case is also referred to as extended spatial scalability (ESS). For quality scalability, the required granularity varies for different applications. In SVC, coarse-grain quality scalability (CGS), which supports bit rate adaptation on a level of coded video sequences, and medium-grain quality scalability (MGS), which supports bit rate adaptation on a NAL unit level, are differentiated.

The standardization activity on SVC started off at the 58th MPEG meeting with an exploration ad hoc group on "inter-frame wavelet video coding" in December 2001 [5]. The work of this ad hoc group and its successor for "Scalable Video Coding" induced a call for proposals for a new standardization activity in October 2003 [6], which resulted in the Scalable Video Model 1.0 [7] in March 2004 that summarizes promising concepts of the submitted proposals. After an evaluation period, the SVC extension of H.264/AVC as proposed in [8] was selected as the starting point for the SVC project in MPEG. In January 2005, this standardization activity became a work item of the Joint Video Team (JVT) of ITU-T SG16/Q6 and ISO/IEC JTC1 SC29/WG11 [9].

In this paper, the key elements of SVC are briefly described and their impact on the coding efficiency for scalable coding is discussed. The rate-distortion performance of SVC for progressive video content is assessed for quality and spatial scalability scenarios as well as for ESS. For evaluating the efficiency of scalable coding with SVC, its coding efficiency is compared to that of state-of-the-art single layer coding with H.264/AVC. Simulation results for MPEG-4 Visual using the Advanced Simple Profile (ASP) are additionally provided to indicate the outstanding rate-distortion performance of both SVC as well as single-layer H.264/AVC coding.

The paper is organized as follows. In the next section, a brief summary of the SVC coder structure is provided. Section III highlights encoder control aspects, while Section IV presents techniques for the extraction of substreams from a given SVC bit stream. In Section V, methods for the determination of suitable configurations of the JSVM software are discussed. Section VI provides rate-distortion plots that allow to compare the coding efficiency of SVC for various scalability scenarios with single layer coding and a simultaneous transmission (simulcast) of single layer bit streams.

Manuscript received October 9, 2006; revised July 13, 2007. This paper was recommended by Guest Editor T. Wiegand.

M. Wien is with the Institut für Nachrichtentechnik, RWTH Aachen University, 52056 Aachen, Germany (e-mail: wien@ient.rwth-aachen.de).

H. Schwarz is with the Fraunhofer-Institute for Telecommunications, Heinrich-Hertz-Institute, 10587 Berlin, Germany (e-mail: heiko.schwarz@hhi.fraunhofer.de).

T. Oelbaum is with the Lehrstuhl für Datenverarbeitung, Technische Universität München, D-80290 München, Germany (e-mail: oelbaum@tum.de).

Digital Object Identifier 10.1109/TCSVT.2007.905530

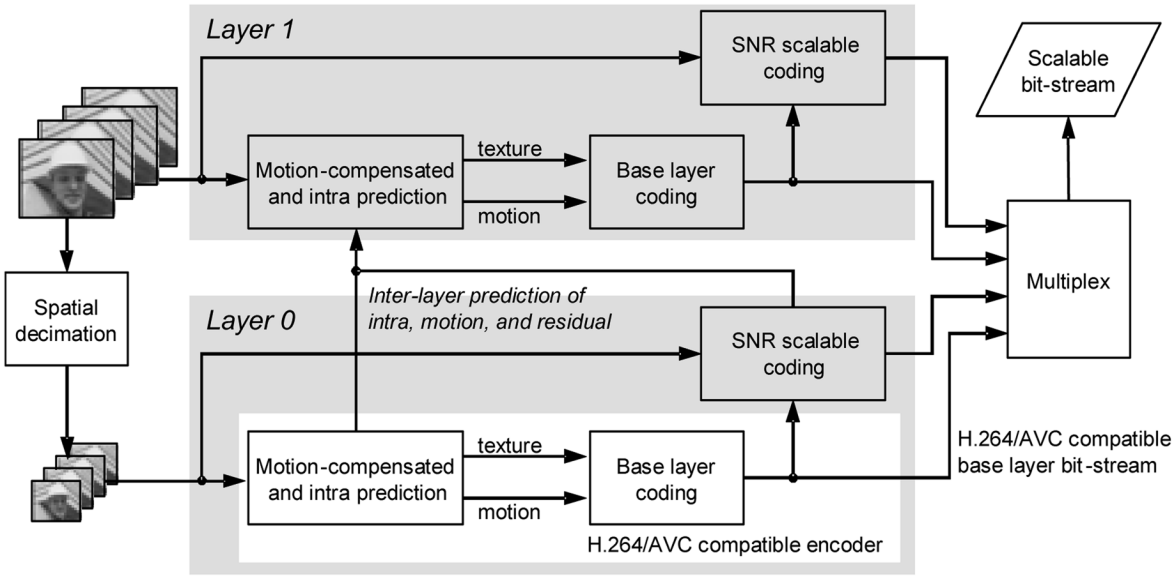

Fig. 1. Coding structure example with two spatial layer.

II. SVC Overview

SVC was designed as an extension of H.264/AVC, and thus most components of H.264/AVC are used as specified in the standard. This includes all key components like motion-compensation, intra prediction, transform and entropy coding, the deblocking filter, or the Network Abstraction Layer (NAL) unit packetization. While temporal scalable coding is already supported in H.264/AVC, new tools are added to enable spatial and quality scalability. Similar to the scalable profiles of the previous video coding standards MPEG-2 Video, H.263, and MPEG-4 Visual, the basic SVC design is mainly determined by the concept for supporting spatial scalable coding and can be classified as a 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.

The supported spatial resolutions are coded in a set of layers which are identified by a dependency identifier D. The base layer has a dependency identifier $D = 0$, and the spatial resolution must not decrease with increasing layer identifier D. For each dependency layer $D > 0$, a reference layer $D_{\mathrm{R}} < D$ can be selected for inter-layer prediction. It is possible to have various layers with identical spatial resolution, but different reconstruction quality. This case is also referred to as CGS. In each spatial or CGS layer, the basic concepts of motion-compensated prediction and intra prediction are employed as in single-layer H.264/AVC. The redundancy between different layers is exploited by additional inter-layer prediction concepts that include prediction mechanisms for macroblock modes and motion parameters as well as texture data (intra and residual). A base quality representation of each layer is obtained by transform coding similar to that of H.264/AVC. The reconstruction quality of this representation can be improved by coding additional quality refinement [signal-to-noise (SNR) refinement] NAL units. These quality refinement NAL units inside a layer are differentiated by a quality identifier Q, which is equal to 0 for the base quality representation and increases with every additional quality refinement representation. In contrast to the layers (D), the spatial resolution must not change between successive quality refinements (Q). However, the prediction mode as well as the applicable motion vectors may be modified between successive quality refinements. While a switching between layers is only specified at defined switching points, switching between different quality representations is possible at any point in time. Hence, a quality scalable bit stream can be flexibly adapted by removing the appropriate quality refinement NAL units. This concept is also referred to as medium-grain quality scalable coding (MGS).

During the development of SVC, another approach for quality scalable coding was investigated, which was based on so-called progressive refinement NAL units. In contrast to all other slice data NAL units, these NAL units could be truncated at any byte-aligned position. Due to its increased computational complexity and the fact that the simple MGS concept already provides a sufficient granularity for quality scalable coding and a similar coding efficiency [10], this approach was finally removed from the SVC specification [11].

Although temporal scalability is already supported in standard H.264/AVC, an additional identifier T for labeling temporal layers is introduced in the SVC high-level syntax. T is equal to 0 for pictures of the temporal base layer and is increased by 1 from one temporal layer to the next.

An important feature of SVC is the provision of scalability at the bit stream level. Bit streams for reduced spatial/temporal resolution and/or bit rate can be simply obtained by discarding NAL units from a global SVC bit stream. Additionally, if only quality scalability is employed and a specific mode of inter-layer prediction is used, the SVC specification enables a lossless and low-complexity rewriting of an SVC stream into a nonscalable H.264/AVC bit stream. The base layer of an SVC bit stream is always coded in compliance with a nonscalable profile of H.264/AVC. In an SVC bit stream, the base layer NAL units are

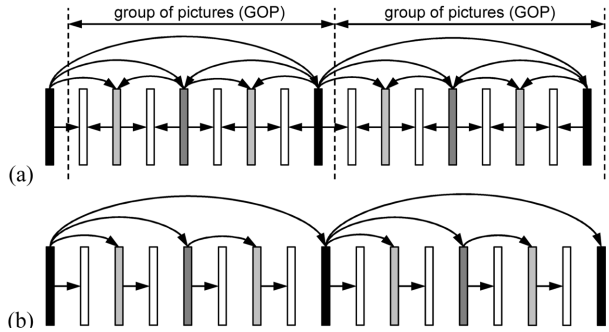

Fig. 2. Hierarchical coding structures for providing temporal scalability. (a) Hierarchical B-pictures. (b) Low-delay coding structure.

prefixed by special SVC NAL units that specify SVC specific parameters for the base layer NAL units including the D, Q, and T values.

The SVC specification defines three scalable profiles. The Scalable Baseline Profile specifies that the base layer must conform to the restricted Baseline Profile and that only restricted spatial scalability configurations are supported. With the Scalable High Profile, a High Profile compliant base layer and fully ESS are supported. The Scalable High Intra Profile includes the same tools as Scalable High, but is restricted to the coding of IDR pictures.

In the following, a brief overview of the basic concepts in SVC for supporting temporal, spatial, and quality scalable coding is given. For more detailed information the reader is referred to the draft standard [11] and the overview in [12].

A. Temporal Scalability

H.264/AVC provides high flexibility in the assignment of reference pictures for motion compensated prediction. Temporal scalable coding can be efficiently provided by using hierarchical coding structures with B- or P-pictures [13], [14] as illustrated in Fig. 2. The pictures of the temporal base layer are only predicted from previous pictures of this layer. The enhancement layer pictures can be bidirectionally predicted by using the two surrounding pictures of a lower temporal layer as references. A picture of the temporal base layer and all temporal refinement pictures between the base layer picture and the previous base layer picture build a group of pictures (GOP).

In addition to enabling temporal scalability, the hierarchical prediction structures usually also provide an improved coding efficiency compared to classical IBBP coding. The coding delay of the hierarchical structures can be controlled by restricting the motion-compensated prediction from pictures of the future. As an example, Fig. 2(b) shows a hierarchical coding structure that provides the same degree of temporal scalability as the one in Fig. 2(a), but with a structural delay of 0. Furthermore, hierarchical prediction structures are not restricted to dyadic temporal scalability, and they can also be combined with the multiple reference picture concept of H.264/AVC. In general, the GOP size or even the prediction structure can be varied over time if intended, e.g., in order to increase the coding efficiency. However, this might restrict the degree of temporal scalability supported by the stream.

B. Spatial and Coarse-Grain Quality Scalability

As illustrated in Fig. 1, spatial scalability is achieved by using a multilayer approach. The pictures of different spatial layers are coded with layer-specific prediction information and motion parameters. In order to improve the enhancement layer coding efficiency in comparison to simulcast, switchable inter-layer prediction mechanisms have been introduced. An encoder can freely choose which information of the reference layer is exploited for efficient enhancement layer coding.

With SVC, each layer can be decoded with a single motion-compensation loop. The employed inter-layer prediction ensures that the computationally complex operations of motion-compensated prediction and deblocking (with the exception of intra-coded blocks) only have to be applied in the target layer, which corresponds to the output pictures. With the exception of intra-coded macroblock that are used for inter-layer prediction, decoded samples of lower layers do not need to be reconstructed. The single-loop decoding feature of SVC is especially important for the case of successive layers of equal spatial resolution (CGS/MGS).

Similar to MPEG-2 Video and MPEG-4 Visual, SVC supports spatial scalability with arbitrary resolution ratios. It is also possible that the enhancement layer only consists of a selected part of the reference layer picture at higher spatial resolution, or that in the enhancement layer additional parts beyond the borders of the reference layer picture are added. This cropping can even be modified on a picture basis.

1) Inter-Layer Motion Prediction: In order to employ motion data from a lower layer for the coding of spatial enhancement layers, a new macroblock type is introduced. This macroblock type is also referred to as reference layer skip mode and it specifies that the prediction data are completely derived from the reference layer and that only a refinement of the residual signal is encoded. When the derived prediction mode specifies inter-picture coding, the macroblock partitioning is determined by up-sampling and realigning the partitioning of the reference layer region that covers the same picture area as the macroblock to be coded. For the simple example of dyadic spatial scalability without cropping, each enhancement layer macroblock corresponds to an 8×8 submacroblock in the reference layer, and thus the enhancement layer macroblock partitioning is obtained by scaling the partitioning of the 8×8 base layer block by a factor of 2 in both vertical and horizontal directions.

In addition to this new macroblock mode, SVC allows to switch between the usual spatial motion vector predictor and an inter-layer motion vector predictor for conventional motion-compensated macroblock coding types. The choice is signaled by a flag that is transmitted on a macroblock partition basis. When inter-layer motion vector prediction is used, the reference frame indexes for the macroblock partition are not transmitted, but derived from the reference layer.

2) Inter-Layer Residual Prediction: The usage of inter-layer residual prediction is signaled by a flag that is transmitted on a macroblock basis. When this flag is true, the corresponding reference layer residual signal is up-sampled and used as a prediction for the residual signal of the current macroblock, so that only the corresponding difference signal is coded. The up-sampling of the reference layer residual is done on a transform block

basis in order to ensure that no filtering across transform block boundaries is applied, which could induce visually disturbing signal components. When the spatial resolution is not modified relative to the reference layer, residual prediction is performed in the transform domain instead of the spatial domain.

3) Inter-Layer Intra Prediction: When a macroblock is coded using the reference layer skip mode as described in Section II-B.1 and the derived prediction mode specifies intra-picture coding, the prediction signal is generated by up-sampling the co-located reconstructed intra signal of the reference layer. This inter-layer intra prediction is the only prediction mode for spatial scalable coding that was already supported in the previous standards that supported spatial scalable coding [1]–[3].

To prevent complete decoding of the lower layers in SVC, the inter-layer intra prediction was restricted to those enhancement layer macroblocks, for which the complete co-located base layer signal is intra-coded. Additionally, constrained intra prediction has to be used in the reference layer such that no inter-predicted samples are employed for intra prediction in the reference layer. With these mandatory restrictions in SVC, each supported layer can be decoded with a single motion compensation loop. The complexity overhead that is required for spatial scalable coding in SVC in comparison to single-layer coding is smaller than in the scalable profiles of MPEG-2 Video, H.263, or MPEG-4 Visual.

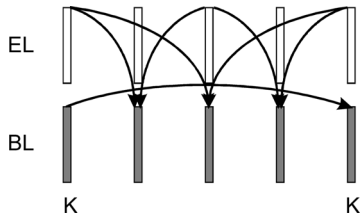

Fig. 3. Key picture concept for hierarchical prediction structures. Between two key pictures (K), the enhancement representation (EL) is used for prediction. For the key pictures, the base representation (BL) is used.

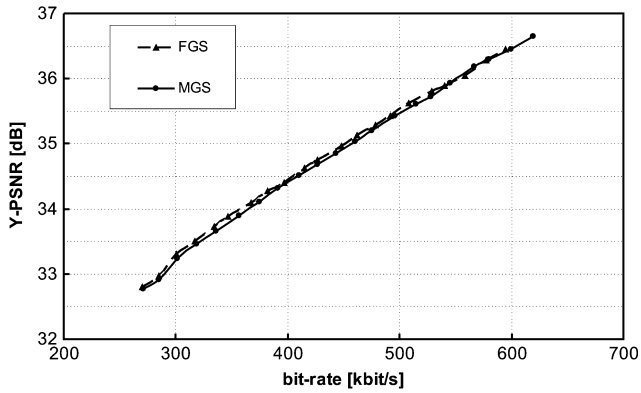

Fig. 4. Comparison of FGS and MGS rate-distortion performance.

C. Medium-Grain Quality Scalability

The quality scalability based on the dependency identifier D as described in Section II-B (CGS) provides only limited granularity, since switching between CGS layers (and spatial layers) is only supported at IDR pictures. In order to increase the granularity for quality scalable coding, SVC provides the possibility to use the quality identifier Q for quality refinements. This method is referred to as medium-grain quality scalability and allows bit stream adaptation an a NAL unit basis; but it requires a concept for controlling the associated drift.

The main reason for the relatively low performance of the fine-granular quality scalability in MPEG-4 Visual is that the motion-compensated prediction is only done in the base layer. In the quality scalable mode of MPEG-2 Video, however, motion-compensated prediction is always employed using the enhancement layer reconstruction as reference. This ensures low-complexity and a high coding efficiency for the enhancement layer, but introduces significant drift when enhancement layer information gets lost.

The SVC design includes a concept for reasonably adjusting the well-known tradeoff between drift and enhancement layer coding efficiency. The concept is designed for quality scalable coding in connection with hierarchical prediction structures. For each picture a flag is transmitted, which signals whether the base representations (when available) or the enhancement representation of the reference pictures are employed for motion-compensated prediction. Pictures that only use the base representations $(Q = 0)$ for prediction are also referred as key pictures. A further flag indicates whether the base representation of a picture is stored in the decoded picture buffer. The key picture concept can be efficiently combined with hierarchical prediction structures as illustrated in Fig. 3. The pictures of the coarsest temporal resolution are transmitted as key pictures, and since only the base representations are used for motion-compensated prediction, no drift occurs between GOPs when any of the enhancement layer NAL units is discarded. The key pictures (K) serve as resynchronization points. For the temporal enhancement pictures however the highest available quality of the reference pictures is used for motion-compensated prediction. This ensures a high coding efficiency, and the drift is efficiently restricted to the inside of a GOP.

As mentioned before, an approach for fine-granular quality scalability (FGS), which enabled byte-aligned NAL unit truncation and used an improved method for entropy coding of transform coefficients in enhancement layers, was investigated during the SVC development [15]. It was, however, shown that the MGS concept is capable of providing very similar rate-distortion results at a significantly reduced complexity, especially for large GOP sizes [10], [16]. In Fig. 4, a representative comparison of FGS and MGS is provided, demonstrating the comparable performance of the two approaches.

III. Encoder Control

A. General Encoder Control

As all video coding standards of ISO/IEC and ITU-T, SVC only specifies the bit stream syntax and the decoding process. However, the encoding process determines the coding efficiency of the generated bit stream to a large extent. Rate-distortion optimized encoder control algorithms based on Lagrangian techniques are well established for single-layer coding due to their simplicity and effectiveness [17], [18]. When coding multiple

layer representations, basically the same concept can be employed. In the encoder control of the Joint Scalable Video Model (JSVM) [19], for each picture, the layers are coded from bottom to top, and in each layer the coding parameters are determined via Lagrangian bit allocation. The additional inter-layer prediction concepts are simply included as additional coding options. This simple concept was used for all experiments in this paper. However, it should be noted that such an encoder control method optimizes the coding parameters from the bottom to the top layer. While the coding efficiency of the base layer is usually identical to that of single-layer coding, the enhancement layer coding efficiency is noticeably worse than that of single-layer coding. Since the coding parameters of the base layer are assumed to be given when coding the enhancement layer, the effective usage of the base layer rate and thus the coding efficiency gain against simulcast is limited. In [20], an encoder control that jointly optimizes the coding parameters for base and enhancement layers is presented. In this paper, it is especially shown that the tradeoff between base and enhancement layer coding efficiency can be adjusted according to the needs of an application.

B. Quantizer Settings for Hierarchical B-Pictures

As reported in the literature [13], [14], the coding efficiency for hierarchical prediction structures is highly dependent on the assignment of quantization parameters to the temporal levels. Intuitively, the pictures of the coarsest temporal resolution should be coded with the highest fidelity, since the reconstruction quality of these pictures influences all other pictures. Following this rule, the quantization step size should be increased with the temporal level identifier T. For the experiments in this paper, the JSVM rule for setting the quantization parameters was used [19], independent of the signal to be coded. With N being the number of temporal layers and a given an initial quantization parameter QP_0, the quantization parameters QP_t for a temporal level $t, 0 \leq t \leq N-1$ are determined by

$$\mathrm{QP}_X = \mathrm{QP}_0 + (t \neq 0\,?\,2:0) - 1.7 \cdot (N-1-t),$$
$$\mathrm{QP}_t = \min(51, \max(0, \mathrm{round}(\mathrm{QP}_X))).$$

This experimentally derived rule provides a QP difference of 1 to 2 between successive temporal levels and an extra refinement for the key pictures.

C. Where to Close the Loop?

For SNR scalable coding with MGS, except for the key pictures, the motion-compensated prediction is always done using the reference with the highest available quality. However, during encoding it is not known what reference will be available in the decoder. Thus, an encoder has to decide what reference it will use for motion estimation, mode decision, and determination of the residual to be coded, i.e., motion compensation. It is often believed that for efficient scalable coding, an open-loop encoder control is required. In open-loop mode the encoder operates decoder unaware, and the pictures of the original input sequence are used as reference for motion estimation and motion compensation, and therewith the impact of quantization on the reconstruction quality is not considered. For single-layer coding it is well-known that closed-loop coding always improved the coding efficiency, since with an open-loop encoder control, the quantization errors of the pictures that are used as reference for motion-compensated prediction cannot be compensated [21], [22].

When SNR refinement layers are present, choosing the right reference is not simple. Based on experimental investigations [21], we employ a concept, for which the prediction loop is closed at two points at the encoder side: the quality base layer and the highest rate point of the SNR refinement enhancement layer. This two-point encoder control was used for all experiments presented in this paper, it has particularly been chosen as a suitable tradeoff between coding efficiency at the lowest and highest supported rate point. For the motion search and the mode decision process, the encoder always uses the reference which has been reconstructed using all available SNR refinement layers. However, for coding the base layer residual, the motion compensation is done using the base layer reference. For the SNR refinement layers, the enhancement layer reference signal is used for motion estimation and mode decision as well as for determination of the residual signal to be coded.

IV. Bit Stream Extraction

An SVC bit stream provides means to extract substreams with lower temporal or spatial resolution, or with reduced reconstruction quality. The extraction of a specific spatio–temporal resolution corresponding to the dependency identifier D_t and the temporal identifier T_t can be done by the following ordered steps:

1) discard all access units with a temporal identifier T greater than T_t;
2) discard all coded slice NAL units with a dependency identifier D greater than D_t.

The extraction of a particular bit rate for a spatio–temporal resolution usually requires that various quality refinement NAL units with a quality identifier $Q > 0$ are additionally discarded from the bit stream. In the following, two methods for extracting a particular bit rate from an SVC bit stream are presented: a simple extraction method based on quality identifiers and an extraction method providing improved rate-distortion performance by using additional priority information.

A. Simple Extraction of a Target Bit Rate

Let R_t specify the target bit rate for a spatio–temporal resolution (D_t, T_t). All packets that do not belong to the spatio–temporal resolution (D_t, T_t) are discarded as described above. Let R_0 be the base bit rate for this spatio–temporal resolution, i.e., the bit rate that corresponds to the base representation with $Q = 0$ including all lower layers that are required for inter-layer prediction of the spatio–temporal resolution (D_t, T_t). R_0 specifies the minimum extractable bit rate for the spatio–temporal target resolution. If R_0 is greater than R_t, the requested spatio–temporal-rate point cannot be extracted from the given SVC bit stream. Otherwise, the following algorithm can be used.

1) The target rate is modified by $R_t = R_t - R_0$.
2) The quality refinement packets $(Q > 0)$ of the spatio–temporal target resolution (D_t, T_t) are processed in increasing order of their quality identifier Q. And for each quality identifier Q, the quality refinement packets are processed in increasing order of their temporal identifier T. For each

set of quality refinement NAL units characterized by the parameters Q and T, the following applies.

a) Let R_{QT} be the bit rate of the set of quality refinement NAL units with quality identifier Q and temporal identifier T of the spatio–temporal resolution (D_t, T_t).
b) If R_{QT} is less or equal to R_t, the corresponding quality refinement packets are included into the extracted bit stream and the target rate is modified by $R_t = R_t - R_{\mathrm{QT}}$. Otherwise, all nonprocessed quality refinement NAL units are discarded and the extraction process is terminated.

In order to meet to target rate R_t more accurately, an appropriate set of quality refinement packets of the first set (Q, T) for which R_{QT} is greater than R_t can be additionally included in the extracted bit stream until the target rate R_t is met.

B. Bit Stream Extraction Using Priority Information

The bit streams that are obtained using the simple extraction method described above are usually characterized by a suboptimal rate-distortion performance, since the extraction algorithm does not take into account the rate-distortion impact of discarding a NAL unit on the remaining bit stream. While the algorithm naturally provides good rate-distortion performance at full quality layers, i.e., when only complete sets of refinement NAL units with the same quality identifier Q are discarded, it performs suboptimal for intermediate rates. The rate-distortion efficiency of extracted bit streams can be improved when the impact of discarding a NAL unit on the rate-distortion efficiency of the remaining bit stream is taken into account. An approach for optimized bit stream extraction using priority information similar to the quality layers in JPEG-2000 [23] is presented in [24]. In SVC, quality layers can be indicated either by making use of the NAL unit header syntax element priority_id or by indication via a separate supplemental enhancement information (SEI) message. Both signaling methods allow up to 64 quality layers to be present in an SVC stream.

The extraction of a spatio–temporal resolution is similar to the method described in Section IV-A. The remaining bit rate budget is compared to the bit rate of a set of refinement NAL units. But instead of processing the sets of quality refinement NAL units with the same values of Q and T, sets of quality refinement packets with the same value of the priority identifiers P are processed.

The rate-distortion efficiency of the bit streams that are obtained with the priority-based extraction is determined by the algorithm that is used for assigning the priority identifiers to the NAL units of an SVC bit stream. Several algorithms for determining priority values that provide a good rate-distortion efficiency for all extractable bit rates have been proposed [24]–[26].

V. JSVM Configuration Optimization

Since the encoder decision on prediction modes and motion parameters may apply for multiple quality refinement layers, the configuration of the applicable residual quantization parameter denoted as RQP is decoupled from the configuration of the Lagrangian multiplier λ for rate-distortion optimization. For easy configurability, the Langrangian multiplier λ that is employed for motion estimation and mode decision is controlled via a mode quantization parameter, which is denoted as MQP. The Lagrangian multiplier λ and the parameter MQP are connected via the relationship given in [17].

In the following a simple quantizer selection method is presented that provides optimized quantizer configurations for the testing conditions employed in this paper. The method provides constant quantizer settings for the whole sequence, and no adaptation over time is employed. As stated before, the method relies on a bottom-up approach, as the optimization of the settings is performed successively for each layer starting with the base layer.

A. Coarse-Grain Scalability and Spatial Scalability

If no quality refinement slices $(Q > 0)$ are used in a configuration, the encoder is operated similar to single-layer encoding [17]. The only difference of the JSVM encoder control in comparison to single-layer coding using Lagrangian bit allocation techniques is that the mode decision for each layer considers the additional SVC macroblock modes with inter-layer prediction in addition to the regular H.264/AVC modes. The Lagrangian multiplier λ is determined depending on the layer quantizer parameter QP $(\mathrm{MQP} = \mathrm{RQP})$ using the same relationship as for the single-layer case [17].

For each layer, the quantization parameter QP is determined to meet the target bit rate as specified in the testing conditions.

B. Quality Scalability

If a testing scenario employs quality refinement layers, the rate-distortion performance over a range of bit rates has to be considered and the determination of a suitable Lagrangian multiplier or the corresponding mode quantization parameter is not straightforward. In the following, a simple search algorithm is presented which was employed to determine the mode quantization parameter MQP for the quality scalability scenarios: With such a configuration, a defined set of rate points $\{R_i\}$ needs to be supported for a spatial resolution D.

1) Set the minimum bit rate R_t for the spatial resolution D.
2) Using a setting with $\mathrm{MQP} = \mathrm{RQP}$:
   a) find the value of $\mathrm{RQP} = \mathrm{RQP}_0$ such that the minimum achievable rate R_t is met;
   b) measure the PSNR values achieved by this configuration at the defined rate points R_i of the current layer D.
3) For $k = 1 \ldots k_{\max}$:
   a) set $\mathrm{MQP}_k = \mathrm{RQP}_0 - k$ and find the value of RQP such that the required minimum rate R_t is met for the current MQP_i;
   b) measure the PSNR values achieved by this configuration at the defined rate points R_i of the current layer D.
4) Determine the best MQP of the set $\{\mathrm{MQP}_k\}$ by determining which value of MQP provided the maximum PSNR over all considered rate points R_i.
5) Assign the best MQP and the corresponding RQP to the layer D.

For quality scalability with one quality refinement layer, often settings in the range of $\mathrm{MQP} = \mathrm{RQP} - 2$ have shown to provide best results. In case of combined scalability scenarios, this algorithm can be successively applied for all spatial layers.

VI. SVC Performance Evaluation

Results are presented according to the SVC testing conditions as defined in [27]. A comparison of the SVC performance for spatial and quality scalability is provided. SVC is compared to single-layer H.264/AVC at the single rate points as well as regarding simulcast test cases. To demonstrate the performance of SVC compared to former standards, rate-distortion results are also presented for MPEG-4 Visual. The performance of ESS is demonstrated by comparing the rate-distortion performance for several scaling relations to dyadic spatial scalability and a simulcast of the spatial layers.

A. Test Conditions

From the beginning of the SVC standardization activity, test conditions were designed to reflect the needs of a broad range of application scenarios such as mobile communications, broadcast, and archival systems. Two scenarios were established to address the diverging needs.

- Scenario I for broad range scalability is motivated by the requirements of applications such as surveillance, broadcast and storage systems.
- Scenario II for limited range scalability is motivated by the requirements of applications such as streaming and mobile communications.

Scenario I covers three layers of spatial resolution (ranging from QCIF to 4CIF) and three layers of temporal resolution (15–60 fps). Random access points to the stream are required every 1.2 s at maximum. Scenario II provides two layers of spatial scalability and two layers of temporal scalability. In this scenario, no random access to the stream is required.

Table I lists the rate points that are required to be extractable from the encoded streams for quality and for dyadic spatial scalability. For the evaluation of quality scalability, the rate points of each row are required to be extractable from the bit streams. For spatial scalability, the rate point in the columns printed bold have to be included in the corresponding streams. For the evaluation, eight test sequences were selected. The four CIF sequences (Bus, Football, Foreman, Mobile) are already well known and used for a long time in the development of video codecs. The four 4CIF sequences (City, Crew, Harbour, Soccer) originate from HD video productions reflecting the special attributes of HD video. For each sequence, the lowest rate points were adjusted for the single layer case to provide an acceptable visual quality.

Results for three scalability tests are provided here.

- Quality scalability: For each spatial layer the performance of quality refinement layers is evaluated.
- Dyadic spatial scalability: The stream consists of two or three spatial layers with dyadic scaling ratio and no quality refinement layers are included. This test enables the assessment of the spatial scalability coding tools performance.
- ESS: The stream consists of two spatial layers with scaling relations of 2/3, 3/4, and 3/5. In the presented testing scenario, the rate of the base layer is kept constant and only the rate of the spatial enhancement layer varies. This test evaluates the performance of nondyadic spatial scalability.

TABLE I
Tested Bit Rates for the Quality and Spatial Scalability Test. For Quality Scalability, Each Row Specifies the Rate Points of One Stream. For Spatial Scalability the Columns Printed in Bold Specify Rate Points of the Streams Under Test

Sequence	Format	Bit rates (kbit/sec) rate point index:				
		0	1	2	3	4
Bus	QCIF 15Hz	**96**	112	**128**	160	**192**
	CIF 30Hz	**384**	448	**512**	640	**768**
Football	QCIF 15Hz	**192**	224	**256**	320	**384**
	CIF 30Hz	**768**	896	**1024**	1280	**1536**
Foreman	QCIF 15Hz	**48**	56	**64**	80	**96**
	CIF 30Hz	**192**	224	**256**	320	**384**
Mobile	QCIF 15Hz	**64**	80	**96**	112	**128**
	CIF 30Hz	**256**	320	**384**	448	**512**
City	QCIF 15Hz	**64**	80	**96**	112	**128**
	CIF 30Hz	**256**	320	**384**	448	**512**
	4CIF 60Hz	**1024**	1280	**1536**	1792	**2048**
Crew, Harbour, Soccer	QCIF 15Hz	**96**	112	**128**	160	**192**
	CIF 30Hz	**384**	448	**512**	640	**768**
	4CIF 60Hz	**1536**	1792	**2048**	2560	**3072**

The bit streams shall be constructed such that successive extraction of included rate points is enabled. In the quality scalability test, each lower rate point must be extractable from a stream with a higher bit rate. For the spatial scalability tests with three spatial layers, the QCIF layer must be extractable from the CIF sequence as well as the 4CIF sequence.

B. Simulation Results for the Dyadic Test Set

Figs. 5 and 6 show exemplary rate-distortion results for the sequences Crew and Foreman for the spatial and quality scalability testing conditions generated with the JSVM 9.1 software. The results are compared to H.264/AVC single layer coding with hierarchical B-pictures. Further, results for H.264/AVC single layer coding with classical IBBP structure (nonhierarchical B-Frames), and results for MPEG-4 Visual ASP are provided. These reference results have been generated using Lagrangian rate-distortion optimization to enable a fair comparison. For the spatial scalability scenario, additional simulcast curves are shown in the plots for the higher spatial layers. The rate-distortion results for simulcast were calculated by successively adding the single layer bit rate necessary to achieve the PSNR of an SVC rate point at the lower layer to the single layer bit rate necessary to achieve the PSNR of an SVC rate point at the higher layer.

The results for the test sequences Crew and Foreman are provided as representative examples for the SVC performance on the full test set. While the absolute numbers change depending on the sequence, the relationship between the different curves remains quite constant for the tested sequences.

It can be seen that the rate-distortion performance for SNR scalability is very close to the rate-distortion performance of the single layer codec. A drop of about 0.5 dB in PSNR or roughly

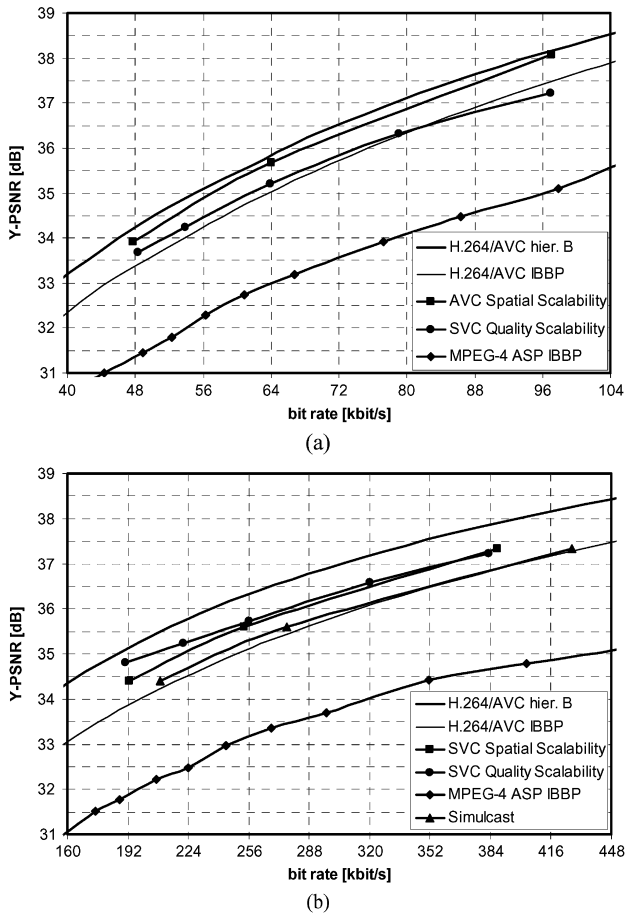

Fig. 5. Test sequence Foreman. (a) QCIF 15 Hz. (b) CIF at 30 Hz. Rate-distortion performance of SVC using quality and spatial scalability to H.264/AVC single layer coding and MPEG-4 ASP. H.264/AVC simulcast for the spatial scalability scenario is additionally shown.

10% rate increase can be observed. Typically, the loss at the lower rate points is lower than the loss at the higher rate points. For spatial scalability, a slight difference between the SVC base layer and the H.264/AVC single layer curve can be observed in Fig. 5(a) and in Fig. 6(a). This performance drop relates to the rate overhead introduced by the SVC high level syntax and the mandatory usage of constrained intra-prediction in SVC layers that are employed for inter-layer prediction. For the CIF and 4CIF resolutions, the spatial scalability curve can be compared to the simulcast of the corresponding H.264/AVC single layer rate points as well as to pure H.264/AVC single layer coding. It can be observed that spatial scalability works especially well for the 4CIF resolution where a PSNR drop of less than 0.5 dB compared to H.264 single layer coding can be observed, and spatial scalability clearly outperforms simulcast.

It has been generally observed that spatial scalability performs better for high resolution input material. For low resolution sequences, the advantage of spatial scalability is smaller as can be seen from the results for the test sequence Foreman.

Generally, it can be seen from the presented plots that H.264/AVC clearly out-performs single layer coding MPEG-4 Visual ASP for the scalable extension, and even more for single layer coding.

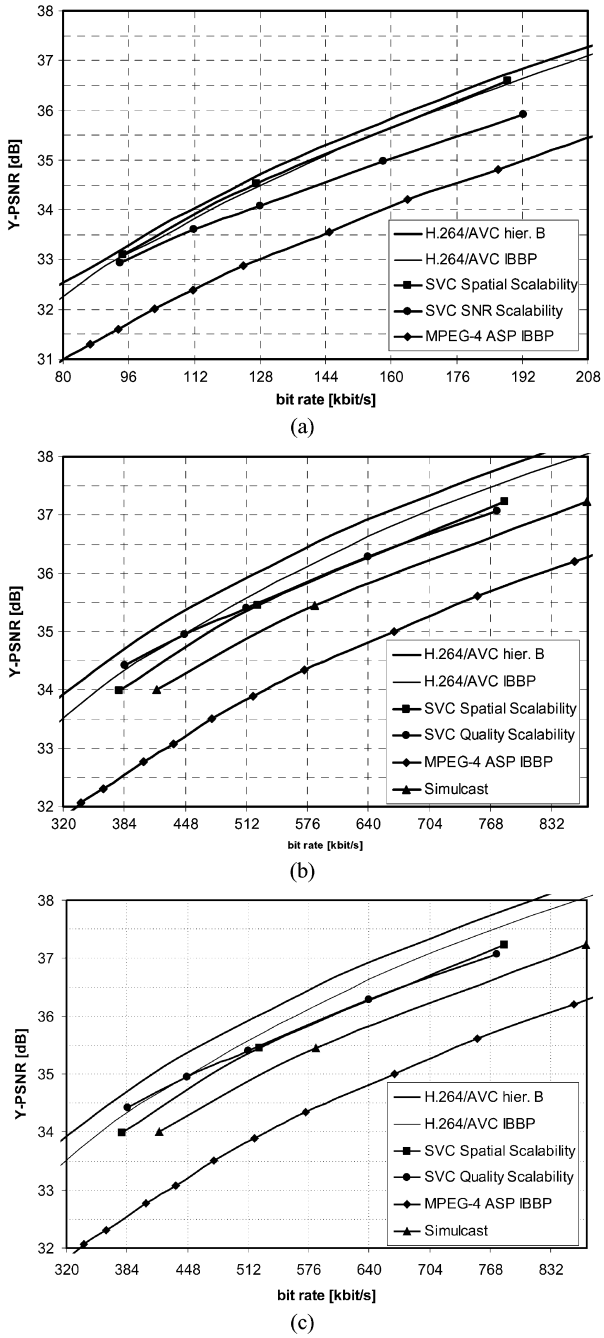

Fig. 6. Test sequence Crew. (a) QCIF 15 Hz. (b) CIF at 30 Hz. (c) 4CIF 60 Hz. Rate-distortion performance of SVC using quality and spatial scalability to H.264/AVC single layer coding and MPEG-4 ASP. In (b), H.264/AVC simulcast for the spatial scalability scenario is additionally shown.

C. Extended Spatial Scalability Performance

Fig. 7 shows exemplary results for the scaling relations 2/3 and 3/4 of the ESS case for the test sequence Crew with 4CIF resolution at the enhancement layer. As discussed in Section II-B, The ESS results are compared to single layer H.264/AVC coding results. To demonstrate the performance of ESS, rate-distortion plots for the simulcast of the fixed low resolution stream with each of the single layer high resolution streams is shown. The curve for the dyadic spatial scalability

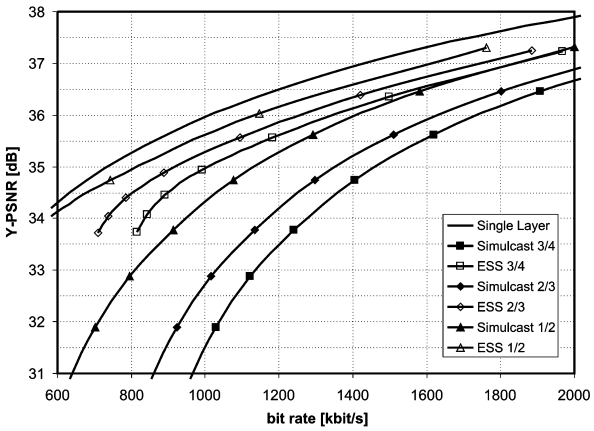

Fig. 7. Test sequence Crew. Results for the ESS test case. The rate of the base layer is as follows: 396 kbps $(1/2)$, 558 kbps $(3/4)$, and 617 kbps $(2/3)$. The resulting PSNR value at the base layer is 36.11 dB for the 1/2 case, 36.94 dB for the 3/4 test case and 36.52 dB for the 2/3 test case.

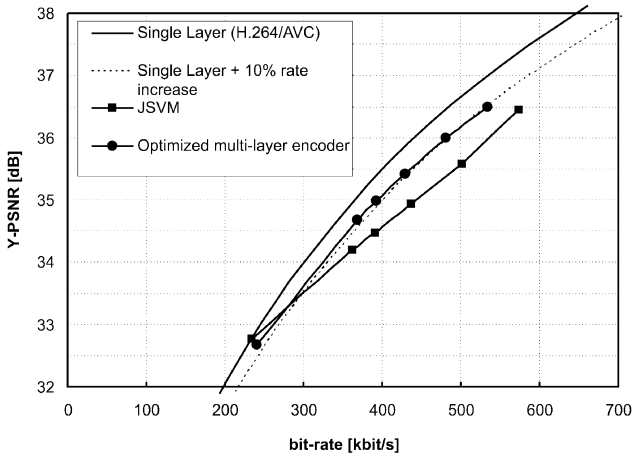

Fig. 8. Rate-distortion efficiency improvement with an optimized multilayer encoder control relative to the JSVM for quality scalable coding of the sequence Soccer in CIF resolution with a frame rate of 30 Hz.

can be seen as a benchmark for the investigated ESS types relative to the dyadic scalability configuration.

The curves reveal that ESS clearly outperforms simulcast coding. At lower rates the differences exceed 2 dB or an equivalent increase in bit rate of more than 50%. As expected the performance loss for ESS compared to dyadic spatial scalability increases when increasing the spatial scalability ratio from 2/3 to a spatial scalability ratio of 3/4.

D. Improved SVC Encoder Control Techniques

As mentioned in Section III-A, the JSVM encoder control [19] specifies a simple bottom-up encoding process. In each spatial or quality enhancement layer, the same Lagrangian bit allocation techniques are employed as for single-layer coding. During the encoding of an enhancement layer, the coding decisions of the reference layers are considered as given. This leads to an uneven distribution of coding efficiency losses between base and enhancement layers. While the base layer rate-distortion performance is virtually identical to that of single-layer coding, significant losses in coding efficiency are typically observed for the enhancement layers. Furthermore, the effective reuse of the reference layer bit rate for enhancement layer coding is limited, because the chosen coding parameters for the reference layer are optimized for that layer only and are not necessarily suitable for an efficient coding of enhancement layers that use the reference layer for inter-layer prediction.

It is possible to improve the coding efficiency of SVC by considering the interdependencies between different layers, which results from the usage of inter-layer prediction, in the encoding control. In [20], first results for spatial and quality scalability by using an improved multilayer encoder control for SVC are presented. During the determination of the coding parameters for layers that are employed for inter-layer prediction of other layers, the impact on the coding efficiency of the dependent layers is taken into account. The achievable improvement using this encoder control technique in comparison to the JSVM encoder control is illustrated in Fig. 8 for the example of quality scalable coding. In comparison to the JSVM encoder control, the optimized encoder control significantly improves the coding efficiency for all supported rate points with exception of the lowest one. Furthermore, it should be noted that quality scalability for the illustrated scenario is provided at the cost of a bit rate increase of only 10% relative to single-layer H.264/AVC coding.

It is expected that further improved encoding techniques for SVC will be developed that take into account the special characteristics of scalable bit streams and improve the SVC coding efficiency in comparison to the simple JSVM encoder control, which was used for all results presented in this paper.

VII. Conclusion

In this paper, the scalable extension of H.264/AVC, SVC, was presented and the performance of the SVC scheme was assessed. The concept of hierarchical B-pictures, which was already included in the first design of H.264/AVC serves as a precursor for SVC. It provides profound performance improvements for both, scalable and single layer coding, and at the same time, provides the—so far missing—basis for efficient SNR scalability. For low-delay applications hierarchical P-pictures can similarly be applied. Although scalable coding still comes at some costs in terms of bit rate (or quality), the gap between H.264/AVC single layer coding and SVC can be remarkably small. Results of the rate-distortion comparison show that SVC clearly outperforms current video coding technologies such as MPEG-4 ASP. Due to the flexibility of the SVC scheme, the layer configuration of a scalable stream can be tailored to the application needs, and thereby, the performance impact of the increased scalability functionality can be controlled. Further research on encoder control for joint optimization of base and enhancement layers is expected to further downscale the gap between single layer coding and scalable coding for dedicated scenarios.

References

[1] *Generic Coding of Moving Pictures and Associated Audio Information—Part 2: Video*, ITU-T and ISO/IEC JTC1, ITU-T Rec. H.262 and ISO/IEC 13818-2 (MPEG-2 Video), Nov. 1994.

[2] *Video Coding for Low Bit Rate Communication*, ITU-T, ITU-T Rec. H.263, Version 1: Nov. 1995, Version 2: Jan. 1998, Version 3: Nov. 2000.
[3] *Coding of Audio-Visual Objects Part 2—Visual*, ISO/IEC JTC1, ISO/IEC 14492-2, (MPEG-4 Visual), Version 1: Apr. 1999, Version 2: Feb. 2000, Version 3: May 2004.
[4] W. Li, "Overview of fine granularity scalability in MPEG-4 video standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 301–317, Mar. 2001.
[5] *Resolutions of 58th WG 11 Meeting*, ISO/IEC JTC 1/SC 29/WG 11, Doc. N4370, Dec. 2001.
[6] *Call for Proposals for Scalable Video Coding*, ISO/IEC JTC 1/SC 29/WG 11, Doc. N5958, Oct. 2003, Brisbane, Australia.
[7] *Scalable Video Model Version 1.0*, ISO/IEC JTC 1/SC 29/WG 11, Doc. N6372, Mar. 2004.
[8] H. Schwarz, T. Hinz, H. Kirchhoffer, D. Marpe, and T. Wiegand, *Technical Description of the HHI Proposal for SVC CE1,* ISO/IEC JTC 1/SC 29/WG 11, Doc. M11244, Oct. 2004.
[9] *Resolutions of 71st WG 11 Meeting*, ISO/IEC JTC 1/SC 29/WG 11, Doc. N6870, Jan. 2005.
[10] H. Schwarz and T. Wiegand, *Implementation and performance of FGS, MGS, and CGS*, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, Doc. JVT-V126, Jan. 2007.
[11] *Joint Draft 11: Scalable Video Coding*, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, Doc. JVT-X201, Jul. 2007.
[12] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable extension of the H.264/MPEG-4 AVC Video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.
[13] H. Schwarz, D. Marpe, and T. Wiegand, *Hierarchical B Pictures*, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, Doc. JVT-P014, Jul. 2005.
[14] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MCTF," in *Proc. ICME*, Toronto, ON, Canada, Jul. 2006, pp. 1929–1932.
[15] *Joint Scalable Video Model*, ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, Doc. JVT-T202, Jul. 2006.
[16] I. Amonou, N. Cammas, S. Kervadec, and S. Pateux, *Enhanced SNR Scalability for Layered CGS Coding Using Quality LAyers*, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, Doc. JVT-S044, Apr. 2006.
[17] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constraint coder control and comparison of video coding standards," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 668–703, Jul. 2003.
[18] K.-P. Lim, *Text Description of Joint Model Reference Encoding Methods and Decoding Concealment Methods*, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, Doc. JVT-L046, Jul. 2004.
[19] *Joint Scalable Video Model*, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, Doc. JVT-X202, Jul. 2007.
[20] H. Schwarz and T. Wiegand, "R-D optimized multilayer encoder control for SVC," presented at the ICIP, San Antonio, TX, Sep. 2007.
[21] H. Schwarz, D. Marpe, and T. Wiegand, *Comparison of MCTF and Closed-Loop Hierarchical B Pictures*, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, Doc. JVT-P059, Jul. 2005.
[22] G. J. Sullivan, *General Characteristics and Design Considerations for Temporal Subband Video Coding*, ITU-T Video Coding Experts Group, Doc. VCEG-U06, Dec. 2003.
[23] D. Taubman and M. Marcellin*, JPEG2000-Image Compression Fundamentals, Standards and Practice*. Norwell, MA: Kluwer, 2002.
[24] I. Amonou, N. Cammas, S. Kervadec, and S. Pateux, "Optimized rate-distortion extraction with quality layers," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1186–1193, Sep. 2007.
[25] H. Schwarz, D. Marpe, and T. Wiegand, *Closed Loop Coding With Quality Layers,* Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, Doc. JVT-Q030, Oct. 2005.
[26] T. Rusert and J.-R. Ohm, "Backward drift estimation with application to quality layer assignment in H.264/AVC based scalable video coding," in *Proc. IEEE ICASSP*, Honolulu, HI, Apr. 2007, pp. 653–656.
[27] M. Wien and H. Schwarz, *Testing Conditions for SVC COding Efficiency and JSVM Performance Evaluation,* Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, Doc. JVT-Q205, Oct. 2005.

Mathias Wien (S'98–M'03) received the diploma and Dr.-Ing. degrees from RWTH Aachen University, Aachen, Germany, in 1997 and 2004, respectively.

From 1997 to 2004, he worked towards the Ph.D. as a Researcher at the Institute of Communications Engineering at RWTH Aachen University, where he is now employed as Senior Research scientist and Head of Administration. His research interests are in the area of image and video processing, space-frequency adaptive and scalable video compression, and robust video transmission. He was an active contributor to the first version of H.264/AVC. He is a Co-Editor of the SVC amendment to H.264/AVC. He is an active contributor to the ITU-T VCEG and the Joint Video Team of VCEG and ISO/IEC MPEG where he co-chaired several AdHoc Groups.

Heiko Schwarz received the Dipl.-Ing. degree in electrical engineering in 1996 and the Dr.-Ing. degree in 2000, both from the University of Rostock, Rostock, Germany.

In 1999, he joined the Fraunhofer Institute for Telecommunications—Heinrich Hertz Institute (HHI), Berlin, Germany. Since then, he has contributed successfully to the standardization activities of the ITU-T Video Coding Experts Group (ITU-T SG16/Q.6—VCEG) and the ISO/IEC Moving Pictures Experts Group (ISO/IEC JTC 1/SC 29/WG 11—MPEG). During the development of the Scalable Video Coding (SVC) extension of H.264/AVC, he co-chaired several ad hoc groups of the Joint Video Team of ITU-T VCEG and ISO/IEC MPEG investigating particular aspects of the scalable video coding design. He has also been appointed as a Co-Editor of the SVC Amendment for H.264/AVC.

Tobias Oelbaum received the Dipl.Ing. degree in electrical engineering and information technology from the Technische Universität München, München, Germany, in 2001, where he is currently working toward the Ph.D. degree.

His research interests are subjective and objective assessment of video quality. He became an active member of MPEG in 2002 and was appointed to chair the test subgroup of MPEG in 2004.