Overview of the Scalable Video Coding Extension of the H.264/AVC Standard

Heiko Schwarz, Detlev Marpe, *Member, IEEE*, and Thomas Wiegand, *Member, IEEE*

***Abstract*—With the introduction of the H.264/AVC video coding standard, significant improvements have recently been demonstrated in video compression capability. The Joint Video Team of the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG) has now also standardized a scalable video coding (SVC) extension of the H.264/AVC standard. SVC enables the transmission and decoding of partial bit streams to provide video services with lower temporal or spatial resolutions or reduced fidelity while retaining a reconstruction quality that is high relative to the rate of the partial bit streams. Hence, SVC provides functionalities such as graceful degradation in lossy transmission environments as well as bit rate, format, and power adaptation. These functionalities provide enhancements to transmission and storage applications. SVC has achieved significant improvements in coding efficiency with an increased degree of supported scalability relative to the scalable profiles of prior video coding standards. This paper provides an overview of the basic concepts for extending H.264/AVC towards SVC. Moreover, the basic tools for providing temporal, spatial, and quality scalability are described in detail and experimentally analyzed regarding their efficiency and complexity.**

***Index Terms*—SVC, H.264, MPEG-4, AVC, standards, video**

I. Introduction

ADVANCES in video coding technology and standardization [1]-[6] along with the rapid developments and improvements of network infrastructures, storage capacity, and computing power are enabling an increasing number of video applications. Application areas today range from multimedia messaging, video telephony, and video conferencing over mobile TV, wireless and wired Internet video streaming, standard- and high-definition TV broadcasting to DVD, Blu-ray Disc, and HD DVD optical storage media. For these applications, a variety of video transmission and storage systems may be employed.

Traditional digital video transmission and storage systems are based on H.222.0 | MPEG-2 systems [7] for broadcasting services over satellite, cable, and terrestrial transmission channels, and for DVD storage, or on H.320 [8] for conversational video conferencing services. These channels are typically characterized by a fixed spatio-temporal format of the video signal (SDTV or HDTV or CIF for H.320 video telephone). Their application behavior in such systems typically falls into one of the two categories: it works or it doesn't work.

Modern video transmission and storage systems using the Internet and mobile networks are typically based on RTP/IP [9] for real-time services (conversational and streaming) and on computer file formats like mp4 or 3gp. Most RTP/IP access networks are typically characterized by a wide range of connection qualities and receiving devices. The varying connection quality is resulting from adaptive resource sharing mechanisms of these networks addressing the time varying data throughput requirements of a varying number of users. The variety of devices with different capabilities ranging from cell phones with small screens and restricted processing power to high-end PCs with high-definition displays results from the continuous evolution of these endpoints.

Scalable video coding (SVC) is a highly attractive solution to the problems posed by the characteristics of modern video transmission systems. The term "scalability" in this paper refers to the removal of parts of the video bit stream in order to adapt it to the various needs or preferences of end users as well as to varying terminal capabilities or network conditions. The term SVC is used interchangeably in this paper for both the concept of scalable video coding in general and for the particular new design that has been standardized as an extension of the H.264/AVC standard. The objective of the SVC standardization has been to enable the encoding of a high-quality video bit stream that contains one or more subset bit streams that can themselves be decoded with a complexity and reconstruction quality similar to that achieved using the existing H.264/AVC design with the same quantity of data as in the subset bit stream.

Scalable video coding has been an active research and standardization area for at least 20 years. The prior international video coding standards H.262 | MPEG-2 Video [3], H.263 [4], and MPEG-4 Visual [5] already include several tools by which the most important scalability modes can be supported. However, the scalable profiles of those standards have rarely been used. Reasons for that include the characteristics of traditional video transmission systems as well as the fact that the spatial and quality scalability features came along with a significant loss in coding efficiency as well as a large increase in decoder complexity as compared to the corresponding non-scalable profiles. It should be noted that two or more single-layer streams, i.e., non-scalable streams, can always be transmitted by the method of *simulcast*, which in principle provides similar functionalities as a scalable bit stream, although typically at

Manuscript received October 6, 2006; revised July 15, 2007.

The authors are with the Fraunhofer Institute for Telecommunications – Heinrich Hertz Institute (HHI), Einsteinufer 37, 10587 Berlin, Germany (e-mail: hschwarz@hhi.fhg.de, marpe@hhi.fhg.de, wiegand@hhi.fhg.de).

the cost of a significant increase in bit rate. Moreover, the adaptation of a single stream can be achieved through transcoding, which is currently used in multipoint control units in video conferencing systems or for streaming services in 3G systems. Hence, a scalable video codec has to compete against these alternatives.

This paper describes the SVC extension of H.264/AVC and is organized as follows. Sec. II explains the fundamental scalability types and discusses some representative applications of scalable video coding as well as their implications in terms of essential requirements. Sec. III gives the history of SVC. Sec. IV briefly reviews basic design concepts of H.264/AVC. In sec. V, the concepts for extending H.264/AVC towards a scalable video coding standard are described in detail and analyzed regarding effectiveness and complexity. The SVC design is summarized in sec. VI. For more detailed information about SVC, the reader is referred to the draft standard [10].

II. Types of Scalability, Applications, and Requirements

In general, a video bit stream is called scalable when parts of the stream can be removed in a way that the resulting sub-stream forms another valid bit stream for some target decoder, and the sub-stream represents the source content with a reconstruction quality that is less than that of the complete original bit stream but is high when considering the lower quantity of remaining data. Bit streams that do not provide this property are referred to as single-layer bit streams. The usual modes of scalability are temporal, spatial, and quality scalability. Spatial scalability and temporal scalability describe cases in which subsets of the bit stream represent the source content with a reduced picture size (spatial resolution) or frame rate (temporal resolution), respectively. With quality scalability, the sub-stream provides the same spatio-temporal resolution as the complete bit stream, but with a lower fidelity – where fidelity is often informally referred to as signal-to-noise ratio (SNR). Quality scalability is also commonly referred to as fidelity or SNR scalability. More rarely required scalability modes are region-of-interest (ROI) and object-based scalability, in which the sub-streams typically represent spatially contiguous regions of the original picture area. The different types of scalability can also be combined, so that a multitude of representations with different spatio-temporal resolutions and bit rates can be supported within a single scalable bit stream.

Efficient scalable video coding provides a number of benefits in terms of applications [11]-[13] – a few of which will be briefly discussed in the following. Consider, for instance, the scenario of a video transmission service with heterogeneous clients, where multiple bit streams of the same source content differing in coded picture size, frame rate, and bit rate should be provided simultaneously. With the application of a properly configured scalable video coding scheme, the source content has to be encoded only once – for the highest required resolution and bit rate, resulting in a scalable bit stream from which representations with lower resolution and/or quality can be obtained by discarding selected data. For instance, a client with restricted resources (display resolution, processing power, or battery power) needs to decode only a part of the delivered bit stream. Similarly, in a multicast scenario, terminals with different capabilities can be served by a single scalable bit stream. In an alternative scenario, an existing video format (like QVGA) can be extended in a backward compatible way by an enhancement video format (like VGA).

Another benefit of scalable video coding is that a scalable bit stream usually contains parts with different importance in terms of decoded video quality. This property in conjunction with unequal error protection is especially useful in any transmission scenario with unpredictable throughput variations and/or relatively high packet loss rates. By using a stronger protection of the more important information, error resilience with graceful degradation can be achieved up to a certain degree of transmission errors. Media-aware network elements (MANEs), which receive feedback messages about the terminal capabilities and/or channel conditions, can remove the non-required parts from a scalable bit stream, before forwarding it. Thus, the loss of important transmission units due to congestion can be avoided and the overall error robustness of the video transmission service can be substantially improved.

Scalable video coding is also highly desirable for surveillance applications, in which video sources not only need to be viewed on multiple devices ranging from high-definition monitors to videophones or PDAs, but also need to be stored and archived. With scalable video coding, for instance, high-resolution/high-quality parts of a bit stream can ordinarily be deleted after some expiration time, so that only low-quality copies of the video are kept for long-term archival. The latter approach may also become an interesting feature in personal video recorders and home networking.

Even though scalable video coding schemes offer such a variety of valuable functionalities, the scalable profiles of existing standards have rarely been used in the past, mainly because spatial and quality scalability have historically come at the price of increased decoder complexity and significantly decreased coding efficiency. In contrast to that, temporal scalability is often supported, e.g., in H.264/AVC-based applications, but mainly because it comes along with a substantial coding efficiency improvement (cp. sec. V.A.2).

H.264/AVC is the most recent international video coding standard. It provides significantly improved coding efficiency in comparison to all prior standards [14]. H.264/AVC has attracted a lot of attention from industry and has been adopted by various application standards and is increasingly used in a broad variety of applications. It is expected that in the near-term future H.264/AVC will be commonly used in most video applications. Given this high degree of adoption and deployment of the new standard and taking into account the large investments that have already been taken place for preparing and developing H.264/AVC-based products, it is quite natural to now build a scalable video coding scheme as an extension of H.264/AVC and to re-use its key features.

Considering the needs of today's and future video applications as well as the experiences with scalable profiles in the past, the success of any future scalable video coding standard critically depends on the following essential requirements:

- Similar coding efficiency compared to single-layer coding – for each subset of the scalable bit stream
- Little increase in decoding complexity compared to single-layer decoding that scales with the decoded spatio-temporal resolution and bit rate
- Support of temporal, spatial, and quality scalability
- Support of a backward compatible base layer (H.264/AVC in this case)
- Support of simple bit stream adaptations after encoding

In any case, the coding efficiency of scalable coding should be clearly superior to that of "simulcasting" the supported spatio-temporal resolutions and bit rates in separate bit streams. In comparison to single-layer coding, bit rate increases of 10% to 50% for the same fidelity might be tolerable depending on the specific needs of an application and the supported degree of scalability.

This paper provides an overview how these requirements have been addressed in the design of the SVC extension of H.264/AVC.

III. History of SVC

Hybrid video coding, as found in H.264/AVC [6] and all past video coding designs that are in widespread application use, is based on motion-compensated temporal differential pulse code modulation (DPCM) together with spatial decorrelating transformations [15]. DPCM is characterized by the use of synchronous prediction loops at the encoder and decoder. Differences between these prediction loops lead to a "drift" that can accumulate over time and produce annoying artifacts. However, the scalability bit stream adaptation operation, i.e., the removal of parts of the video bit stream can produce such differences.

Subband or transform coding does not have the drift property of DPCM. Therefore, video coding techniques based on motion-compensated 3-d wavelet transforms have been studied extensively for use in scalable video coding [16]-[19]. The progress in wavelet-based video coding caused MPEG to start an activity on exploring this technology. As a result, MPEG issued a call for proposals for efficient scalable video coding technology in October 2003 with the intention to develop a new scalable video coding standard. 12 of the 14 submitted proposals in response to this call [20] represented scalable video codecs based on 3-d wavelet transforms, while the remaining two proposals were extensions of H.264/AVC [6]. After a 6 month evaluation phase, in which several subjective tests for a variety of conditions were carried out and the proposals were carefully analyzed regarding their potential for a successful future standard, the scalable extension of H.264/AVC as proposed in [21] was chosen as the starting point [22] of MPEG's scalable video coding (SVC) project in October 2004. In January 2005, MPEG and VCEG agreed to jointly finalize the SVC project as an Amendment of H.264/AVC within the Joint Video Team.

Although the initial design [21] included a wavelet-like decomposition structure in temporal direction, it was later removed from the SVC specification [10]. Reasons for that removal included drastically reduced encoder and decoder complexity and improvements in coding efficiency. It was shown that an adjustment of the DPCM prediction structure can lead to a significantly improved drift control as will be shown in the paper. Despite this change, most components of the proposal in [21] remained unchanged from the first model [22] to the latest draft [10] being augmented by methods for non-dyadic scalability and interlaced processing which were not included in the initial design.

IV. H.264/AVC Basics

SVC was standardized as an extension of H.264/AVC. In order to keep the paper self-contained, the following brief description of H.264/AVC is limited to those key features that are relevant for understanding the concepts of extending H.264/AVC towards scalable video coding. For more detailed information about H.264/AVC, the reader is referred to the standard [6] or corresponding overview papers [23]-[26].

Conceptually, the design of H.264/AVC covers a *Video Coding Layer* (VCL) and a *Network Abstraction Layer* (NAL). While the VCL creates a coded representation of the source content, the NAL formats these data and provides header information in a way that enables simple and effective customization of the use of VCL data for a broad variety of systems.

A. Network Abstraction Layer (NAL)

The coded video data are organized into NAL units, which are packets that each contains an integer number of bytes. A NAL unit starts with a one-byte header, which signals the type of the contained data. The remaining bytes represent payload data. NAL units are classified into VCL NAL units, which contain coded slices or coded slice data partitions, and non-VCL NAL units, which contain associated additional information. The most important non-VCL NAL units are parameter sets and supplemental enhancement information (SEI). The sequence and picture parameter sets contain infrequently changing information for a video sequence. SEI messages are not required for decoding the samples of a video sequence. They provide additional information which can assist the decoding process or related processes like bit stream manipulation or display. A set of consecutive NAL units with specific properties is referred to as an access unit. The decoding of an access unit results in exactly one decoded picture. A set of consecutive access units with certain properties is referred to as a coded video sequence. A coded video sequence represents an independently decodable part of a NAL unit bit stream. It always starts with an instantaneous decoding refresh (IDR) access unit, which signals that the IDR access unit and all following access units can be decoded without decoding any previous pictures of the bit stream.

B. Video Coding Layer (VCL)

The VCL of H.264/AVC follows the so-called block-based hybrid video coding approach. Although its basic design is very similar to that of prior video coding standards such as H.261, MPEG-1 Video, H.262 | MPEG-2 Video, H.263, or MPEG-4 Visual, H.264/AVC includes new features that enable it to achieve a significant improvement in compression efficiency relative to any prior video coding standard [14]. The main difference to previous standards is the largely increased flexibility and adaptability of H.264/AVC.

The way pictures are partitioned into smaller coding units in H.264/AVC, however, follows the rather traditional concept of subdivision into *macroblocks* and *slices*. Each picture is partitioned into macroblocks that each covers a rectangular picture area of 16×16 luma samples and, in the case of video in 4:2:0 chroma sampling format, 8×8 samples of each of the two chroma components. The samples of a macroblock are either spatially or temporally predicted, and the resulting prediction residual signal is represented using transform coding. The macroblocks of a picture are organized in slices, each of which can be parsed independently of other slices in a picture. Depending on the degree of freedom for generating the prediction signal, H.264/AVC supports three basic slice coding types:

- *I* slice: *intra-picture* predictive coding using spatial prediction from neighboring regions,
- *P* slice: intra-picture predictive coding and inter-picture *predictive* coding with one prediction signal for each predicted region,
- *B* slice: intra-picture predictive coding, inter-picture predictive coding, and inter-picture *bi-predictive* coding with two prediction signals that are combined with a weighted average to form the region prediction.

For *I* slices, H.264/AVC provides several directional spatial intra prediction modes, in which the prediction signal is generated by using neighboring samples of blocks that precede the block to be predicted in coding order. For the luma component, the intra prediction is either applied to 4×4, 8×8, or 16×16 blocks, whereas for the chroma components, it is always applied on a macroblock basis[1].

For *P* and *B* slices, H.264/AVC additionally permits variable block size motion-compensated prediction with multiple reference pictures [27]. The macroblock type signals the partitioning of a macroblock into blocks of 16×16, 16×8, 8×16, or 8×8 luma samples. When a macroblock type specifies partitioning into four 8×8 blocks, each of these so-called *sub-macroblocks* can be further split into 8×4, 4×8, or 4×4 blocks, which is indicated through the sub-macroblock type. For *P* slices, one motion vector is transmitted for each block. In addition, the used reference picture can be independently chosen for each 16×16, 16×8, or 8×16 macroblock partition or 8×8 sub-macroblock. It is signaled via a reference index parameter, which is an index into a list of reference pictures that is replicated at the decoder.

In *B* slices, two distinct reference picture lists are utilized, and for each 16×16, 16×8, or 8×16 macroblock partition or 8×8 sub-macroblock, the prediction method can be selected between *list 0*, *list 1*, or *bi-prediction*. While list 0 and list 1 prediction refer to unidirectional prediction using a reference picture of reference picture list 0 or 1, respectively, in the bi-predictive mode, the prediction signal is formed by a weighted sum of a list 0 and list 1 prediction signal. In addition, special modes as so-called *direct modes* in *B* slices and *skip modes* in *P* and *B* slices are provided, in which such data as motion vectors and reference indices are derived from previously transmitted information.

For transform coding, H.264/AVC specifies a set of *integer transforms* of different block sizes. While for intra macroblocks the transform size is directly coupled to the intra prediction block size, the luma signal of motion-compensated macroblocks that do not contain blocks smaller than 8×8 can be coded by using either a 4×4 or 8×8 transform. For the chroma components a two-stage transform, consisting of 4×4 transforms and a Hadamard transform of the resulting DC coefficients is employed[1]. A similar hierarchical transform is also used for the luma component of macroblocks coded in intra 16×16 mode. All inverse transforms are specified by exact integer operations, so that inverse-transform mismatches are avoided. H.264/AVC uses *uniform reconstruction quantizers*. One of 52 quantization step sizes[1] can be selected for each macroblock by the quantization parameter *QP*. The scaling operations for the quantization step sizes are arranged with logarithmic step size increments, such that an increment of the *QP* by 6 corresponds to a doubling of quantization step size.

For reducing blocking artifacts, which are typically the most disturbing artifacts in block-based coding, H.264/AVC specifies an *adaptive deblocking filter*, which operates within the motion-compensated prediction loop.

H.264/AVC supports two methods of entropy coding, which both use context-based adaptivity to improve performance relative to prior standards. While *CAVLC* (context-based adaptive variable-length coding) uses variable-length codes and its adaptivity is restricted to the coding of transform coefficient levels, *CABAC* (context-based adaptive binary arithmetic coding) utilizes arithmetic coding and a more sophisticated mechanism for employing statistical dependencies, which leads to typical bit rate savings of 10-15% relative to CAVLC.

In addition to the increased flexibility on the macroblock level, H.264/AVC also allows much more flexibility on a picture and sequence level compared to prior video coding standards. Here we mainly refer to *reference picture memory control*. In H.264/AVC, the coding and display order of pictures is completely decoupled. Furthermore, any picture can be marked as reference picture for use in motion-compensated prediction of following pictures, independent of the slice coding types. The behavior of the *decoded picture buffer* (DPB), which can hold up to 16 frames (depending on the used con-

[1] Some details of the profiles of H.264/AVC that were designed primarily to serve the needs of professional application environments are neglected in this description, particularly in relation to chroma processing and range of step sizes.

formance point and picture size), can be adaptively controlled by *memory management control operation* (MMCO) commands, and the reference picture lists that are used for coding of *P* or *B* slices can be arbitrarily constructed from the pictures available in the DPB via *reference picture list re-ordering* (RPLR) commands.

In order to enable a flexible partitioning of a picture into slices, the concept of *slice groups* was introduced in H.264/AVC. The macroblocks of a picture can be arbitrarily partitioned into slice groups via a *slice group map*. The slice group map, which is specified by the content of the picture parameter set and some slice header information, assigns a unique slice group identifier to each macroblock of a picture. And each slice is obtained by scanning the macroblocks of a picture that have the same slice group identifier as the first macroblock of the slice in raster-scan order. Similar to prior video coding standards, a *picture* comprises the set of slices representing a complete frame or one field of a frame (such that, e.g., an interlaced-scan picture can be either coded as a single frame picture or two separate field pictures). Additionally, H.264/AVC supports a macroblock-adaptive switching between frame and field coding. For that, a pair of vertically adjacent macroblocks is considered as a single coding unit, which can be either transmitted as two spatially-neighboring frame macroblocks, or as interleaved top and a bottom field macroblocks.

V. Basic Concepts for Extending H.264/AVC Toward a Scalable Video Coding Standard

Apart from the required support of all common types of scalability, the most important design criteria for a successful scalable video coding standard are coding efficiency and complexity, as was noted in sec. II. Since SVC was developed as an extension of H.264/AVC with all of its well-designed core coding tools being inherited, one of the design principles of SVC was that new tools should only be added if necessary for efficiently supporting the required types of scalability.

A. Temporal scalability

A bit stream provides temporal scalability when the set of corresponding access units can be partitioned into a temporal base layer and one or more temporal enhancement layers with the following property. Let the temporal layers be identified by a temporal layer identifier *T*, which starts from 0 for the base layer and is increased by 1 from one temporal layer to the next. Then for each natural number *k*, the bit stream that is obtained by removing all access units of all temporal layers with a temporal layer identifier *T* greater than *k* forms another valid bit stream for the given decoder.

For hybrid video codecs, temporal scalability can generally be enabled by restricting motion-compensated prediction to reference pictures with a temporal layer identifier that is less than or equal to the temporal layer identifier of the picture to be predicted. The prior video coding standards MPEG-1 [2], H.262 | MPEG-2 Video [3], H.263 [4], and MPEG-4 Visual [5] all support temporal scalability to some degree. H.264/AVC [6] provides a significantly increased flexibility for temporal scalability because of its reference picture memory control. It allows the coding of picture sequences with arbitrary temporal dependencies, which are only restricted by the maximum usable DPB size. Hence, for supporting temporal scalability with a reasonable number of temporal layers, no changes to the design of H.264/AVC were required. The only related change in SVC refers to the signaling of temporal layers, which is described in sec. VI.

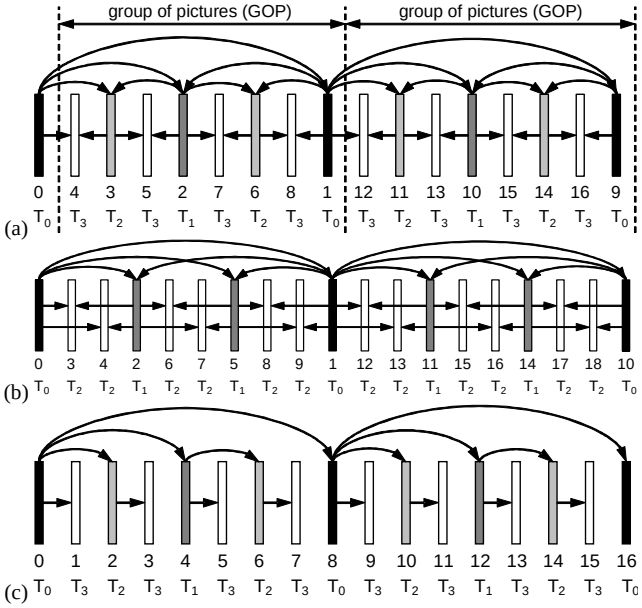


Fig. 1. Hierarchical prediction structures for enabling temporal scalability: (a) coding with hierarchical B pictures, (b) non-dyadic hierarchical prediction structure, (c) hierarchical prediction structure with a structural encoder/decoder delay of zero. The numbers directly below the pictures specify the coding order, the symbols T_k specify the temporal layers with *k* representing the corresponding temporal layer identifier.

1) Hierarchical prediction structures

Temporal scalability with dyadic temporal enhancement layers can be very efficiently provided with the concept of hierarchical B pictures [28][29] as illustrated in Fig. 1a[2]. The enhancement layer pictures are typically coded as *B* pictures, where the reference picture lists 0 and 1 are restricted to the temporally preceding and succeeding picture, respectively, with a temporal layer identifier less than the temporal layer identifier of the predicted picture. Each set of temporal layers $\{T_0,\ldots,T_k\}$ can be decoded independently of all layers with a temporal layer identifier $T > k$. In the following, the set of pictures between two successive pictures of the temporal base layer together with the succeeding base layer picture is referred to as a *group of pictures* (GOP).

Although the described prediction structure with hierarchical *B* pictures provides temporal scalability and also shows excellent coding efficiency as will be demonstrated later, it

[2] As described above, neither P or B slices are directly coupled with the management of reference pictures in H.264/AVC. Hence, backward prediction is not necessarily coupled with the use of B slices and the temporal coding structure of Fig. 1a can also be realized using P slices resulting in a structure that is often called hierarchical P pictures.

represents a special case. In general, hierarchical prediction structures for enabling temporal scalability can always be combined with the multiple reference picture concept of H.264/AVC. This means that the reference picture lists can be constructed by using more than one reference picture, and they can also include pictures with the same temporal level as the picture to be predicted. Furthermore, hierarchical prediction structures are not restricted to the dyadic case. As an example, Fig. 1b illustrates a non-dyadic hierarchical prediction structure, which provides 2 independently decodable sub-sequences with 1/9-th and 1/3-rd of the full frame rate. It should further be noted that it is possible to arbitrarily modify the prediction structure of the temporal base layer, e.g., in order to increase the coding efficiency. The chosen temporal prediction structure does not need to be constant over time.

Note that it is possible to arbitrarily adjust the structural delay between encoding and decoding a picture by restricting motion-compensated prediction from pictures that follow the picture to be predicted in display order. As an example, Fig. 1c shows a hierarchical prediction structure, which does not employ motion-compensated prediction from pictures in the future. Although this structure provides the same degree of temporal scalability as the prediction structure of Fig. 1a, its structural delay is equal to zero compared to 7 pictures for the prediction structure in Fig. 1a. However, such low-delay structures typically decrease coding efficiency.

The coding order for hierarchical prediction structures has to be chosen in a way that reference pictures are coded before they are employed for motion-compensated prediction. This can be ensured by different strategies, which mostly differ in the associated decoding delay and memory requirement. For a detailed analysis the reader is referred to [28][29].

The coding efficiency for hierarchical prediction structures is highly dependent on how the quantization parameters are chosen for pictures of different temporal layers. Intuitively, the pictures of the temporal base layer should be coded with highest fidelity, since they are directly or indirectly used as references for motion-compensated prediction of pictures of all temporal layers. For the next temporal layer a larger quantization parameter should be chosen, since the quality of these pictures influences fewer pictures. Following this rule, the quantization parameter should be increased for each subsequent hierarchy level. Additionally, the optimal quantization parameter also depends on the local signal characteristics.

An improved selection of the quantization parameters can be achieved by a computationally expensive rate-distortion analysis similar to the strategy presented in [30]. In order to avoid such a complex operation, we have chosen the following strategy (cp. [31]), which proved to be sufficiently robust for a wide range of tested sequences. Based on a given quantization parameter QP_0 for pictures of the temporal base layer, the quantization parameters for enhancement layer pictures of a given temporal layer with an identifier $T > 0$ are determined by $QP_T = QP_0 + 3 + T$. Although this strategy for cascading the quantization parameters over hierarchy levels results in relatively large PSNR fluctuations inside a group of pictures, subjectively, the reconstructed video appears to be temporally smooth without annoying temporal "pumping" artifacts.

Often, motion vectors for bi-predicted blocks are determined by independent motion searches for both reference lists. It is, however, well-known that the coding efficiency for *B* slices can be improved when the combined prediction signal (weighted sum of list 0 and list 1 predictions) is considered during the motion search, e.g. by employing the iterative algorithm presented in [32].

When using hierarchical B pictures with more than 2 temporal layers, it is also recommended to use the "spatial direct mode" of the H.264/AVC inter-picture prediction design [6], since with the "temporal direct mode" unsuitable "direct motion vectors" are derived for about half of the B pictures. It is also possible to select between the spatial and temporal direct mode on a picture basis.

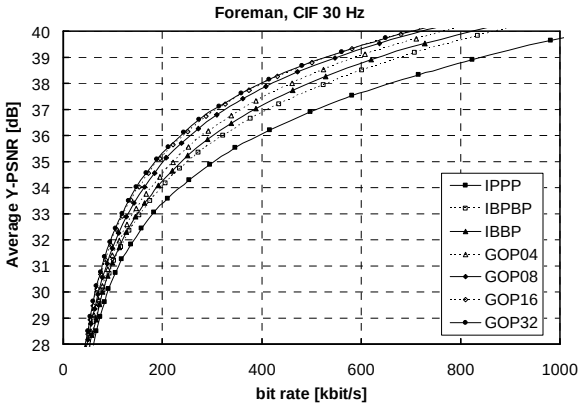


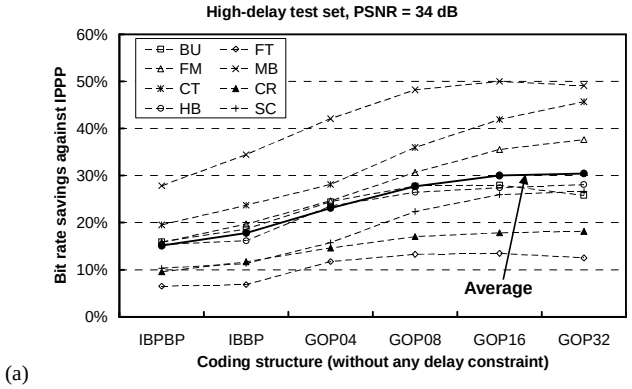
Fig. 2. Coding efficiency comparison of hierarchical B pictures without any delay constraints and conventional IPPP, IBPBP, and IBBP coding structures for the sequence "Foreman" in CIF resolution and a frame rate of 30 Hz.

2) Coding efficiency of hierarchical prediction structures

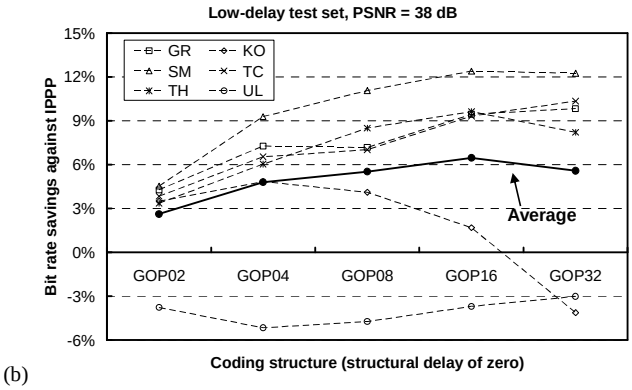
We now analyze the coding efficiency of dyadic hierarchical prediction structures for both high- and low-delay coding. The encodings were operated according to the Joint Scalable Video Model (JSVM) algorithm [31]. The sequences were encoded using the High Profile of H.264/AVC, and CABAC was selected as entropy coding method. The number of active reference pictures in each list was set to 1 picture.

In a first experiment we analyze coding efficiency for hierarchical B pictures without applying any delay constraint. Fig. 2 shows a representative result for the sequence "Foreman" in CIF (352×288) resolution and a frame rate of 30 Hz. The coding efficiency can be continuously improved by enlarging the GOP size up to about 1 second. In comparison to the widely used IBBP coding structure, PSNR gains of more than 1 dB can be obtained for medium bit rates in this way. For the sequences of the high-delay test set (see TABLE I) in CIF resolution and a frame rate of 30 Hz, the bit rate savings at an acceptable video quality of 34 dB that are obtained by using hierarchical prediction structures in comparison to IPPP coding are summarized in Fig. 3a. For all test sequences, the coding efficiency can be improved by increasing the GOP size and

thus the encoding/decoding delay; the maximum coding efficiency is achieved for GOP sizes between 8 and 32 pictures.



(a)



(b)

Fig. 3. Bit rate savings for various hierarchical prediction structures relative to IPPP coding: (a) Simulations without any delay constraint for the high-delay test set (see TABLE I), (b) Simulations with a structural delay of zero for the low-delay test set (see TABLE II).

In a further experiment the structural encoding/decoding delay is constrained to be equal to zero and the coding efficiency of hierarchical prediction structures is analyzed for the video conferencing sequences of the low-delay test set with a resolution of 368×288 samples and with a frame rate of 25 Hz or 30 Hz. The bit rate savings in comparison to IPPP coding, which is commonly used in low-delay applications, for an acceptable video quality of 38 dB are summarized in Fig. 3b. In comparison to hierarchical coding without any delay constraint the coding efficiency improvements are significantly smaller. However, for most of the sequences we still observe coding efficiency gains relative to IPPP coding. From these experiments, it can be deduced that providing temporal scalability usually doesn't have any negative impact on coding efficiency. Minor losses in coding efficiency are possible when the application requires low delay. However, especially when a higher delay can be tolerated, the usage of hierarchical prediction structures not only provides temporal scalability, but also significantly improves coding efficiency.

B. Spatial scalability

For supporting spatial scalable coding, SVC follows the conventional approach of multi-layer coding, which is also used in H.262 | MPEG-2 Video, H.263, and MPEG-4 Visual. Each layer corresponds to a supported spatial resolution and is referred to by a spatial layer or *dependency identifier D*. The dependency identifier *D* for the base layer is equal to 0, and it is increased by 1 from one spatial layer to the next. In each spatial layer, motion-compensated prediction and intra prediction are employed as for single-layer coding. But in order to improve coding efficiency in comparison to simulcasting different spatial resolutions, additional so-called *inter-layer prediction* mechanisms are incorporated as illustrated in Fig. 4.

In order to restrict the memory requirements and decoder complexity, SVC specifies that the same coding order is used for all supported spatial layers. The representations with different spatial resolutions for a given time instant form an access unit and have to be transmitted successively in increasing order of their corresponding spatial layer identifiers *D*. But as illustrated in Fig. 4, lower layer pictures do not need to be present in all access units, which makes it possible to combine temporal and spatial scalability.

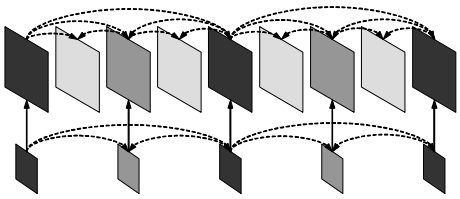


Fig. 4. Multi-layer structure with additional inter-layer prediction for enabling spatial scalable coding.

1) Inter-layer prediction

The main goal when designing inter-layer prediction tools is to enable the usage of as much lower layer information as possible for improving rate-distortion efficiency of the enhancement layers. In H.262 | MPEG-2 Video, H.263, and MPEG-4 Visual, the only supported inter-layer prediction methods employ the reconstructed samples of the lower layer signal. The prediction signal is either formed by motion-compensated prediction inside the enhancement layer, by upsampling the reconstructed lower layer signal, or by averaging such an upsampled signal with a temporal prediction signal.

Although the reconstructed lower layer samples represent the complete lower layer information, they are not necessarily the most suitable data that can be used for inter-layer prediction. Usually, the inter-layer predictor has to compete with the temporal predictor, and especially for sequences with slow motion and high spatial detail, the temporal prediction signal mostly represents a better approximation of the original signal than the upsampled lower layer reconstruction. In order to improve the coding efficiency for spatial scalable coding, two additional inter-layer prediction concepts [33] have been added in SVC: *prediction of macroblock modes and associated motion parameters* and *prediction of the residual signal.*

When neglecting the minor syntax overhead for spatial enhancement layers, the coding efficiency of spatial scalable coding should never become worse than that of simulcast, since in SVC, all inter-layer prediction mechanisms are switchable. An SVC conforming encoder can freely choose between intra- and inter-layer prediction based on the given local signal characteristics. Inter-layer prediction can only take place inside a given access unit using a layer with a spatial

layer identifier *D* less than the spatial layer identifier of the layer to be predicted. The layer that is employed for inter-layer prediction is also referred to as *reference layer*, and it is signaled in the slice header of the enhancement layer slices. Since the SVC inter-layer prediction concepts include techniques for motion as well as residual prediction, an encoder should align the temporal prediction structures of all spatial layers.

Although the SVC design supports spatial scalability with arbitrary resolution ratios [34][35], for the sake of simplicity, we restrict our following description of the inter-layer prediction techniques to the case of dyadic spatial scalability, which is characterized by a doubling of the picture width and height from one layer to the next. Extensions of these concepts will be briefly summarized in sec. V.B.2.

a) Inter-layer motion prediction

For spatial enhancement layers, SVC includes a new macroblock type, which is signaled by a syntax element called *base mode flag*. For this macroblock type, only a residual signal but no additional side information such as intra prediction modes or motion parameters is transmitted. When *base mode flag* is equal to 1 and the corresponding 8×8 block[3] in the reference layer lies inside an intra-coded macroblock, the macroblock is predicted by *inter-layer intra prediction* as will be explained in sec. V.B.1c. When the reference layer macroblock is inter-coded, the enhancement layer macroblock is also inter-coded. In that case, the partitioning data of the enhancement layer macroblock together with the associated reference indices and motion vectors are derived from the corresponding data of the co-located 8×8 block in the reference layer by so-called *inter-layer motion prediction*.

The macroblock partitioning is obtained by upsampling the corresponding partitioning of the co-located 8×8 block in the reference layer. When the co-located 8×8 block is not divided into smaller blocks, the enhancement layer macroblock is also not partitioned. Otherwise, each $M \times N$ sub-macroblock partition in the 8×8 reference layer block corresponds to a $(2M) \times (2N)$ macroblock partition in the enhancement layer macroblock. For the upsampled macroblock partitions, the same reference indices as for the co-located reference layer blocks are used; and both components of the associated motion vectors are derived by scaling the corresponding reference layer motion vector components by a factor of 2.

In addition to this new macroblock type, the SVC concept includes the possibility to use scaled motion vectors of the co-located 8×8 block in the reference layer as motion vector predictors for conventional inter-coded macroblock types. A flag for each used reference picture list that is transmitted on a macroblock partition level, i.e., for each 16×16, 16×8, 8×16, or 8×8 block, indicates whether inter-layer motion vector predictor is used. If this so-called *motion prediction flag* for a reference picture list is equal to 1, the corresponding reference indices for the macroblock partition are not coded in the enhancement layer, but the reference indices of the co-located reference layer macroblock partition are used, and the corresponding motion vector predictors for all blocks of the enhancement layer macroblock partition are formed by the scaled motion vectors of the co-located blocks in the reference layer. A *motion prediction flag* equal to 0 specifies that the reference indices for the corresponding reference picture list are coded in the enhancement layer (when the number of active entries in the reference picture list is greater than 1 as specified by the slice header syntax) and that conventional spatial motion vector prediction as specified in H.264/AVC is employed for the motion vectors of the corresponding reference picture list.

[3] Note that for conventional dyadic spatial scalability, a macroblock in a spatial enhancement layer corresponds to an 8×8 sub-macroblock in its reference layer.

b) Inter-layer residual prediction

Inter-layer residual prediction can be employed for all inter-coded macroblocks regardless whether they are coded using the newly introduced SVC macroblock type signaled by the *base mode flag* or by using any of the conventional macroblock types. A flag is added to the macroblock syntax for spatial enhancement layers, which signals the usage of inter-layer residual prediction. When this *residual prediction flag* is equal to 1, the residual signal of the corresponding 8×8 sub-macroblock in the reference layer is block-wise upsampled using a bi-linear filter and used as prediction for the residual signal of the enhancement layer macroblock, so that only the corresponding difference signal needs to be coded in the enhancement layer. The upsampling of the reference layer residual is done on a transform block basis in order to ensure that no filtering is applied across transform block boundaries, by which disturbing signal components could be generated [36]. Fig. 5 illustrates the visual impact of upsampling the residual by filtering across block boundary and the block-based filtering in SVC.

Fig. 5. Visual example for the enhancement layer when filtering across residual block boundaries (left) and omitting filtering across residual block boundaries (right) for residual prediction.

c) Inter-layer intra prediction

When an enhancement layer macroblock is coded with *base mode flag* equal to 1 and the co-located 8×8 sub-macroblock in its reference layer is intra-coded, the prediction signal of the enhancement layer macroblock is obtained by *inter-layer intra prediction*, for which the corresponding reconstructed intra signal of the reference layer is upsampled. For upsampling the luma component, one-dimensional 4-tap FIR filters are applied horizontally and vertically. The chroma components are up-

sampled by using a simple bi-linear filter. Filtering is always performed across sub-macroblock boundaries using samples of neighboring intra blocks. When the neighboring blocks are not intra-coded, the required samples are generated by specific border extension algorithms. In this way, it is avoided to reconstruct inter-coded macroblocks in the reference layer and thus, so-called *single-loop decoding* is provided [37][38], which will be further explained in sec. V.B.3 below. To prevent disturbing signal components in the prediction signal, the H.264/AVC deblocking filter is applied to the reconstructed intra signal of the reference layer before upsampling.

2) Generalized spatial scalability

Similar to H.262 | MPEG-2 Video and MPEG-4 Visual, SVC supports spatial scalable coding with arbitrary resolution ratios. The only restriction is that neither the horizontal nor the vertical resolution can decrease from one layer to the next. The SVC design further includes the possibility that an enhancement layer picture represents only a selected rectangular area of its corresponding reference layer picture, which is coded with a higher or identical spatial resolution. Alternatively, the enhancement layer picture may contain additional parts beyond the borders of the reference layer picture. This reference and enhancement layer *cropping*, which may also be combined, can even be modified on a picture-by-picture basis.

Furthermore, the SVC design also includes tools for spatial scalable coding of interlaced sources. For both extensions, the generalized spatial scalable coding with arbitrary resolution ratios and cropping as well as for the spatial scalable coding of interlaced sources, the three basic inter-layer prediction concepts are maintained. But especially the derivation process for motion parameters as well as the design of appropriate upsampling filters for residual and intra blocks needed to be generalized. For a detailed description of these extensions the reader is referred to [34] and [35].

It should be noted that in an extreme case of spatial scalable coding, both the reference and the enhancement layer may have the same spatial resolution and the cropping may be aligned with macroblock boundaries. As a specific feature of this configuration, the deblocking of the reference layer intra signal for inter-layer intra prediction is omitted, since the transform block boundaries in the reference layer and the enhancement layer are aligned. Furthermore, inter-layer intra and residual prediction are directly performed in the transform coefficient domain in order to reduce the decoding complexity. When a reference layer macroblock contains at least one non-zero transform coefficient, the co-located enhancement layer macroblock has to use the same luma transform size (4×4 or 8×8) as the reference layer macroblock.

3) Complexity considerations

As already pointed out, the possibility of employing inter-layer intra prediction is restricted to selected enhancement layer macroblocks, although coding efficiency can typically be improved (see sec. V.B.4) by generally allowing this prediction mode in an enhancement layer, as it was done in the initial design [33]. In [21] and [37], however, it was shown that decoder complexity can be significantly reduced by constraining the usage of inter-layer intra prediction. The idea behind this so-called *constrained inter-layer prediction* is to avoid the computationally complex and memory access intensive operations of motion compensation and deblocking for inter-coded macroblocks in the reference layer. Consequently, the usage of inter-layer intra prediction is only allowed for enhancement layer macroblocks, for which the co-located reference layer signal is intra-coded. It is further required that all layers that are used for inter-layer prediction of higher layers are coded using constrained intra prediction, so that the intra-coded macroblocks of the reference layers can be constructed without reconstructing any inter-coded macroblock.

Under these restrictions, which are mandatory in SVC, each supported layer can be decoded with a *single motion compensation loop*. Thus, the overhead in decoder complexity for SVC compared to single-layer coding is smaller than that for prior video coding standards, which all require multiple motion compensation loops at the decoder side. Additionally, it should be mentioned that each quality or spatial enhancement layer NAL unit can be parsed independently of the lower layer NAL units, which provides further opportunities for reducing the complexity of decoder implementations [39].

4) Coding efficiency

The effectiveness of the SVC inter-layer prediction techniques for spatial scalable coding has been evaluated in comparison to single-layer coding and simulcast. For this purpose, the base layer was coded at a fixed bit rate, whereas for encoding the spatial enhancement layer, the bit rate as well as the amount of enabled inter-layer prediction mechanisms was varied. Additional simulations have been performed by allowing an unconstrained inter-layer intra prediction and hence decoding with multiple motion compensation loops. Only the first access unit was intra-coded and CABAC was used as entropy coding method. Simulations have been carried out for a GOP size of 16 pictures as well as for IPPPP coding. All encoders have been rate-distortion optimized according to [14]. For each access unit, first the base layer is encoded, and given the corresponding coding parameters, the enhancement layer is coded [31]. The inter-layer prediction tools are considered as additional coding options for the enhancement layer in the operational encoder control. The lower resolution sequences have been generated following the method in [31]. The simulation results for the sequences "City" and "Crew" with spatial scalability from CIF (352×288) to 4CIF (704×576) and a frame rate of 30 Hz are depicted in Fig. 6. For both sequences, results for a GOP size of 16 pictures (providing 5 temporal layers) are presented while for "Crew", also a result for IPPP coding (GOP size of 1 picture) is depicted. For all cases, all inter-layer prediction (ILP) tools, given as intra (I), motion (M), and residual (R) prediction, improve the coding efficiency in comparison to simulcast. However, the effectiveness of a tool or a combination of tools strongly depends on the sequence characteristics and the prediction structure. While the result for the sequence "Crew" and a GOP size of 16 pictures

is very close to that for single-layer coding, some losses are visible for "City", which is the worst performing sequence in our test set. Moreover, as illustrated for "Crew", the overall performance of SVC compared to single-layer coding reduces when moving from a GOP size of 16 pictures to IPPP coding.

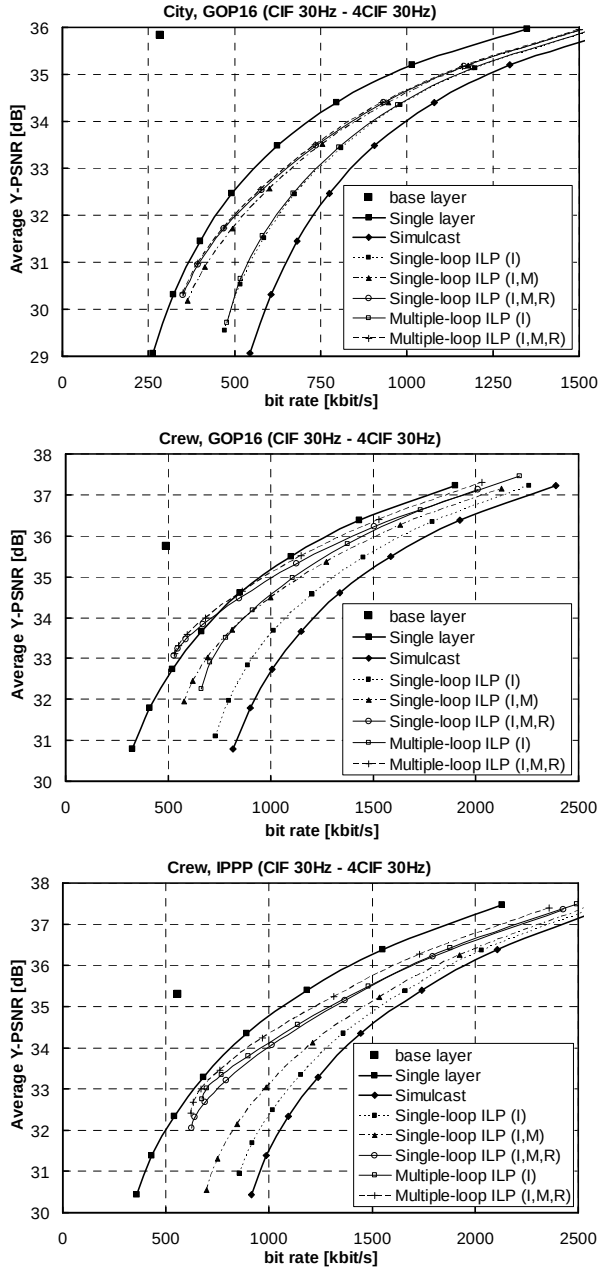


Fig. 6. Efficiency analysis of the inter-layer prediction concepts in SVC for different sequences and prediction structures. The rate-distortion point for the base layer is plotted as a solid rectangle inside the diagrams, but it should be noted that it corresponds to a different spatial resolution.

Multiple-loop decoding can further improve the coding efficiency as illustrated in Fig. 6. But the gain is often minor and comes at the price of a significant increase in decoder complexity. It is worth noting that the rate-distortion performance for multi-loop decoding using only inter-layer intra prediction ("multiple-loop ILP (I)") is usually worse than that of the "single-loop ILP (I,M,R)" case, where the latter corresponds to the fully featured SVC design while the former is conceptually comparable to the scalable profiles of H.262 | MPEG-2 Video, H.263, or MPEG-4 Visual. However, it should be noted that the hierarchical prediction structures which not only improve the overall coding efficiency but also the effectiveness of the inter-layer prediction mechanisms, are not supported in these prior video coding standards.

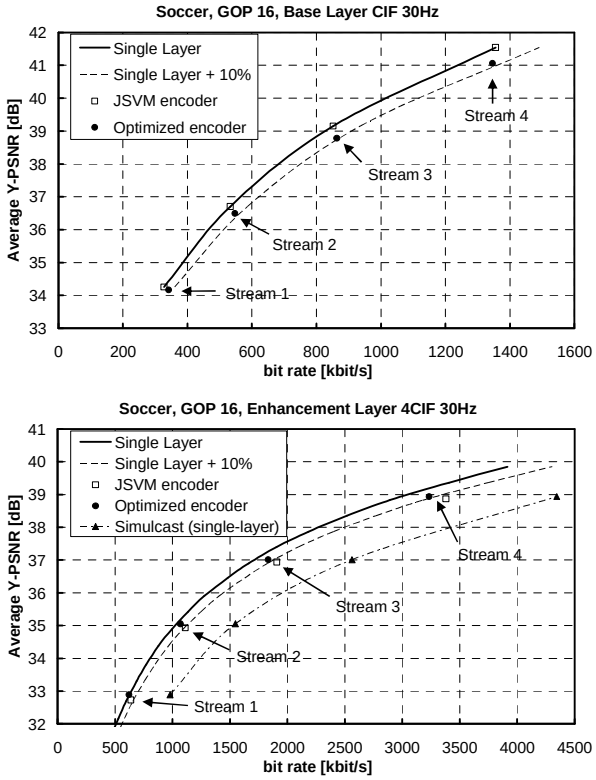


Fig. 7. Experimental results for spatial scalable coding (from CIF to 4CIF, 30 Hz) of the sequence "Soccer" using an optimized encoder control.

5) Encoder control

The encoder control as used in the JSVM [31] for multi-layer coding represents a bottom-up process. For each access unit, first the coding parameters of the base layer are determined, and given these data, the enhancement layers are coded in increasing order of their layer identifier *D*. Hence, the results in Fig. 6 show only losses for the enhancement layer while the base layer performance is identical to that for single-layer H.264/AVC coding. However, this encoder control concept might limit the achievable enhancement layer coding efficiency, since the chosen base layer coding parameters are only optimized for the base layer, but they are not necessarily suitable for an efficient enhancement layer coding. A similar effect might be observed when using different downsampled sequences as input for the base layer coding. While the encoder control for the base layer minimizes the reconstruction error relative to each individual downsampled "original", the different obtained base layer coding parameters may result in more or less re-usable data for the enhancement layer coding, although the reconstructed base layer sequences may have a subjectively comparable reconstruction quality.

First experimental results for an improved multi-layer encoder control which takes into account the impact of the base

layer coding decisions on the rate-distortion efficiency of the enhancement layers are presented in [40]. The algorithm determines the base layer coding parameters using a weighted sum of the Lagrangian costs for base and enhancement layer. Via the corresponding weighting factor it is possible to trade-off base and enhancement layer coding efficiency. In Fig. 7, an example result for spatial scalable coding with hierarchical B pictures and a GOP size of 16 pictures is shown. Four scalable bit streams have been coded with both the JSVM and the optimized encoder control. The quantization parameter QP_E for the enhancement layer was set to $QP_B + 4$, with QP_B being the quantization parameter for the base layer. With the optimized encoder control the SVC coding efficiency can be controlled in a way that the bit rate increase relative to single layer coding for the same fidelity is always less or equal to 10% for both the base and the enhancement layer.

C. Quality scalability

Quality scalability can be considered as a special case of spatial scalability with identical picture sizes for base and enhancement layer. As already mentioned in sec. V.B, this case is supported by the general concept for spatial scalable coding and it is also referred to as *coarse-grain quality scalable coding* (CGS). The same inter-layer prediction mechanisms as for spatial scalable coding are employed, but without using the corresponding upsampling operations and the inter-layer deblocking for intra-coded reference layer macroblocks. Furthermore, the inter-layer intra and residual prediction are directly performed in the transform domain. When utilizing inter-layer prediction for coarse-grain quality scalability in SVC, a refinement of texture information is typically achieved by requantizing the residual texture signal in the enhancement layer with a smaller quantization step size relative to that used for the preceding CGS layer.

However, this multi-layer concept for quality scalable coding only allows a few selected bit rates to be supported in a scalable bit stream. In general, the number of supported rate points is identical to the number of layers. Switching between different CGS layers can only be done at defined points in the bit stream (cp. sec. VI). Furthermore, as will be demonstrated in sec. V.C.4, the multi-layer concept for quality scalable coding becomes less efficient, when the relative rate difference between successive CGS layers gets smaller.

Especially for increasing the flexibility of bit stream adaptation and error robustness, but also for improving the coding efficiency for bit streams that have to provide a variety of bit rates, a variation of the CGS approach, which is also referred to as *medium-grain quality scalability* (MGS), is included in the SVC design. The differences to the CGS concept are a modified high-level signaling (cp. sec. VI), which allows a switching between different MGS layers in any access unit, and the so-called *key picture concept* (cp. sec. V.C.1), which allows the adjustment of a suitable trade-off between drift and enhancement layer coding efficiency for hierarchical prediction structures. With the MGS concept, any enhancement layer NAL unit can be discarded from a quality scalable bit stream, and thus packet-based quality scalable coding is provided. SVC additionally provides the possibility to distribute the enhancement layer transform coefficients among several slices. To this end, the first and the last scan index for transform coefficients are signaled in the slice headers, and the slice data only include transform coefficient levels for scan indices inside the signaled range. Thus, the information for a quality refinement picture that corresponds to a certain quantization steps size can be distributed over several NAL units corresponding to different quality refinement layers with each of them containing refinement coefficients for particular transform basis functions only (cp. [41]). In addition, the macroblocks of a picture (and a quality refinement layer) can be partitioned into several slices as in standard H.264/AVC.

1) Controlling drift in quality scalable coding

The process of motion-compensated prediction for packet-based quality scalable coding has to be carefully designed, since it determines the trade-off between enhancement layer coding efficiency and *drift* (cp. [42]). Drift describes the effect that the motion-compensated prediction loops at encoder and decoder are not synchronized, e.g., because quality refinement packets are discarded from a bit stream. Fig. 8 illustrates different concepts for trading off enhancement layer coding efficiency and drift for packet-based quality scalable coding.

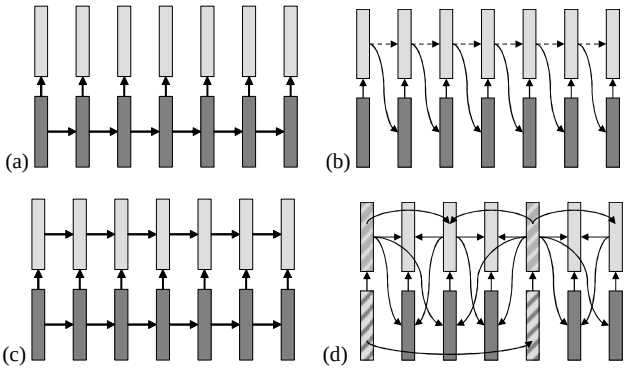


Fig. 8. Various concepts for trading off enhancement layer coding efficiency and drift for packet-based quality scalable coding: (a) base layer only control, (b) enhancement layer only control, (c) two-loop control, (d) key picture concept of SVC for hierarchical prediction structures, where key pictures are marked by the hatched boxes.

For *fine-grain quality scalable* (FGS) coding in MPEG-4 Visual, the prediction structure was chosen in a way that drift is completely omitted. As illustrated in Fig. 8a, motion compensation in MPEG-4 FGS is only performed using the base layer reconstruction as reference, and thus any loss or modification of a quality refinement packet doesn't have any impact on the motion compensation loop. The drawback of this approach, however, is that it significantly decreases enhancement layer coding efficiency in comparison to single-layer coding. Since only base layer reconstruction signals are used for motion-compensated prediction, the portion of bit rate that is spent for encoding MPEG-4 FGS enhancement layers of a picture cannot be exploited for the coding of following pictures that use this picture as reference.

For quality scalable coding in H.262 | MPEG-2 Video, the

other extreme case of possible prediction structures was specified. Here, the reference with the highest available quality is always employed for motion-compensated prediction as depicted in Fig. 8b[4]. This enables highly efficient enhancement layer coding and ensures low complexity, since only a single reference picture needs to be stored for each time instant. However, any loss of quality refinement packets results in a drift[5] that can only be controlled by intra updates.

As an alternative, a concept with two motion compensation loops as illustrated in Fig. 8c could be employed. This concept is similar to spatial scalable coding as specified in H.262 | MPEG-2 Video, H.263, and MPEG-4 Visual. Although the base layer is not influenced by packet losses in the enhancement layer, any loss of a quality refinement packet results in a drift for the enhancement layer reconstruction.

For MGS coding in SVC an alternative approach using so-called *key pictures* [21] has been introduced. For each picture a flag is transmitted, which signals whether the base quality reconstruction or the enhancement layer reconstruction of the reference pictures is employed for motion-compensated prediction. In order to limit the memory requirements, a second syntax element signals whether the base quality representation of a picture is additionally reconstructed and stored in the decoded picture buffer. In order to limit the decoding overhead for such key pictures, SVC specifies that motion parameters must not change between the base and enhancement layer representations of key pictures, and thus also for key pictures, the decoding can be done with a single motion-compensation loop. Fig. 8d illustrates how the key picture concept can be efficiently combined with hierarchical prediction structures.

All pictures of the coarsest temporal layer are transmitted as key pictures, and only for these pictures the base quality reconstruction is inserted in the decoded picture buffer. Thus, no drift is introduced in the motion compensation loop of the coarsest temporal layer. In contrast to that, all temporal refinement pictures typically use the reference with the highest available quality for motion-compensated prediction, which enables a high coding efficiency for these pictures. Since the key pictures serve as re-synchronization points between encoder and decoder reconstruction, drift propagation is efficiently limited to neighboring pictures of higher temporal layers. The trade-off between enhancement layer coding efficiency and drift can be adjusted by the choice of the GOP size or the number of hierarchy stages. It should be noted that both the quality scalability structure in H.262 | MPEG-2 Video (no picture is coded as key picture) and the FGS coding approach in MPEG-4 Visual (all pictures are coded as key pictures) basically represent special cases of the SVC key picture concept.

[4] For a generalization of the basic concept, Fig. 8b indicates (by dashed arrows) that motion parameters may be changed between base and enhancement layer, although this is not supported in H.262 | MPEG-2 Video.

[5] Since H.262 | MPEG-2 Video does not allow partial discarding of quality refinement packets inside a video sequence, the drift issue can be completely avoided in conforming H.262 | MPEG-2 Video bit streams by controlling the reconstruction quality of both the base and the enhancement layer during encoding (cp. sec. V.C.4).

2) Encoder control

As described in the previous section, except for key pictures, motion-compensated prediction for quality scalable coding is always performed by employing the highest available quality of the corresponding reference pictures. However, during the encoding process for MGS layers it is not known what representation will be available in the decoder. The encoder has to decide what reference it will use for motion estimation, mode decision, and the determination of the residual signal to be coded (motion compensation). This decision influences the coding efficiency for the supported rate points. Several investigations [44][45] turned out that a good coding efficiency is usually obtained when the prediction loop in the encoder is closed at the highest rate point, i.e., for the processes of motion estimation, mode decision, and motion compensation the references with the highest reconstruction quality are employed. Note that this is different from so-called open-loop coding where the original of the reference pictures is used. In [44][45] it is additionally pointed out that the coding efficiency of the base layer can be improved by a two-loop encoder control, in which the base layer residual to be coded is determined by a second motion compensation process for which the base layer references are used. The impact on enhancement layer coding efficiency is typically small. In order to further improve the enhancement layer coding efficiency, the optimized encoder control mentioned in sec. V.B.5 can also be employed for quality scalable coding.

3) Bit stream extraction

For extracting a sub-stream with a particular average bit rate from a given quality scalable bit stream (using the MGS approach) usually a huge number of possibilities exist. The same average bit rate can be adjusted by discarding different quality refinement NAL units. Thus, the obtained average reconstruction error that corresponds to the given target bit rate may depend on the used extraction method. A very simple method may consist of randomly discarding MGS refinement packets until the requested bit rate is reached. Alternatively, in a more sophisticated method, a priority identifier is assigned to each coded slice NAL unit by an encoder. During the bit stream extraction process, at first, coded slice NAL units with the lowest priority are discarded, and when the target bit rate is not already reached coded slice NAL units of the next priority class are discarded, etc. The priority identifiers can either be fixed by the encoder based on the employed coder structure or determined by a rate-distortion analysis. The SVC syntax (cp. sec. VI) provides different means for including such priority information in a bit stream. For more detailed information about the concept of optimized bit stream extraction, which is also referred to as *priority layers*, the reader is referred to [46].

4) Coding efficiency

In a first experiment the different concepts for controlling drift, as discussed in sec. V.C.1, are evaluated for hierarchical *B* pictures with a GOP size of 16 pictures. With exception of the 2-loop control, all configurations could be realized with an SVC compliant encoder. Results for the sequences "City" and

"Crew" are summarized in Fig. 9. For these simulations, the intermediate rate points for all drift control concepts were obtained by randomly discarding quality refinement NAL units.

When the motion compensation loop is closed at the base layer (BL-only control) as in MPEG-4 FGS (corresponding to Fig. 8a), no drift occurs, but the enhancement layer coding efficiency is very low, especially for sequences like "City" for which motion-compensated prediction works very well.

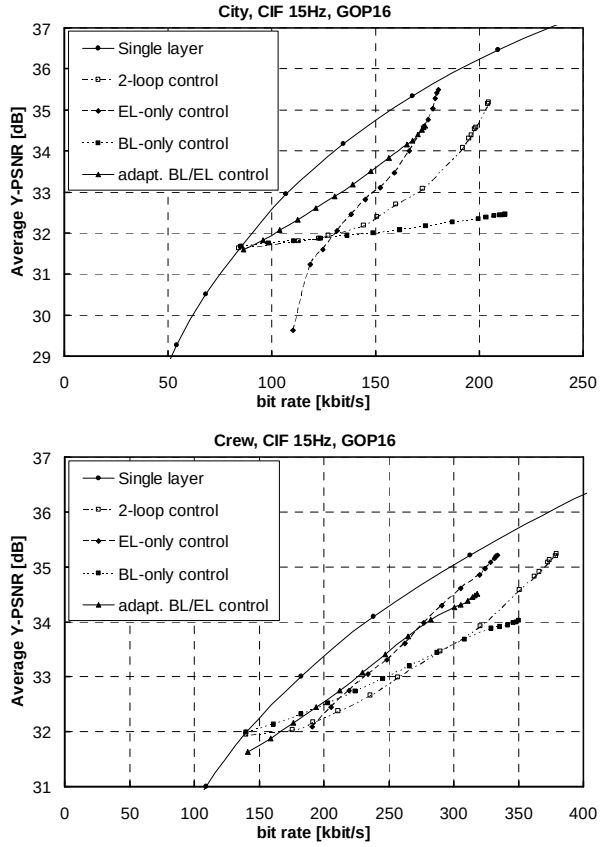


Fig. 9. Comparison of drift control concepts with different tradeoffs between enhancement layer coding efficiency and drift for the sequences "City" and "Crew" in CIF resolution and a frame rate of 15 Hz.

By closing the loop only at the enhancement layer (EL-only control), as it is done in the quality scalable mode of H.262 | MPEG-2 Video (corresponding to Fig. 8b), a high enhancement layer coding efficiency can be achieved. But the discarding of enhancement layer packets typically results in a serious drift, and the reconstructed video quickly becomes unusable. It should be noted that the behavior of the enhancement layer only control highly depends on the employed encoder control concept. For the simulations in Fig. 9, the encoder control was operated with the goal to optimize the enhancement layer coding efficiency. With a different encoder control, it is possible to obtain a base layer that has the same coding efficiency as a single-layer bit stream. However, such an encoder control significantly reduces the enhancement layer coding efficiency. And regardless of the used encoder control, a partial loss of the enhancement layer NAL units always results in a significant drift.

A similar behavior can also be observed for the 2-loop control (corresponding to Fig. 8c), but here the reconstruction quality stabilizes for low rates at the base layer level. For the sequence "Crew" the corresponding impact is less obvious, since a substantial portion of macroblocks is intra-coded and the differences only apply for inter coding.

With the SVC key picture concept (adapt. BL/EL control – corresponding to Fig. 8d), in which the pictures of the coarsest temporal level are coded as key pictures, a reasonable coding efficiency for the entire supported rate interval can be achieved in connection with hierarchical prediction structures. The results in Fig. 9 also show that the SVC design can only provide a suitable coding efficiency for quality scalable coding with a wide range of supported bit rates when hierarchical prediction structures are employed.

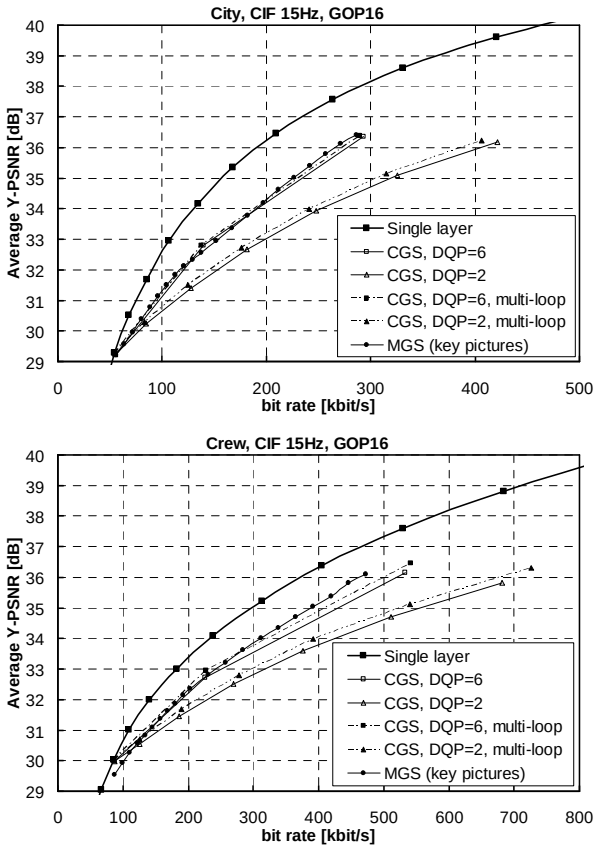


Fig. 10. Comparison of coarse-grain and medium-grain quality scalable coding with different configurations for the sequences "City" and "Crew" in CIF resolution and a frame rate of 15 Hz.

In a second experiment different configurations for providing quality scalability are evaluated. In Fig. 10, the coding efficiency of CGS coding and MGS coding with key pictures is compared to that of single-layer coding for hierarchical B pictures with a GOP size of 16 pictures. For the quality scalable bit streams, the bit rate interval between the lowest and highest supported rate point corresponds to a *QP* difference of 12, i.e., the enhancement layer quantization step is equal to 1/4th of the base layer quantization step size. By comparing different CGS configurations with different choices of *delta QP* (DQP), which is the numerical difference between the *QP* values of two successive layers, it can be seen that coding effi-

ciency generally decreases with an increasing number of supported rate points, i.e., with decreasing DQP. The diagrams also contain rate-distortion curves for CGS with multiple-loop decoding, which is not supported by the SVC design. As already observed for spatial scalable coding, multiple-loop decoding for CGS increases coding efficiency only slightly and therefore, it does not justify the corresponding increase in decoder complexity relative to single-loop decoding. Additionally, Fig. 10 also shows the coding efficiency of the more flexible MGS coding with the usage of the key picture concept and a DQP of 6. The improved coding efficiency at the highest rate point and the reduced coding efficiency at the lowest rate point for the MGS runs in comparison to the CGS runs with DQP equal to 6 are a result of the improved encoder control for MGS, which is described in sec. V.C.2. It should be noted that with MGS coding, the number of supported rate points is significantly increased in comparison to CGS coding.

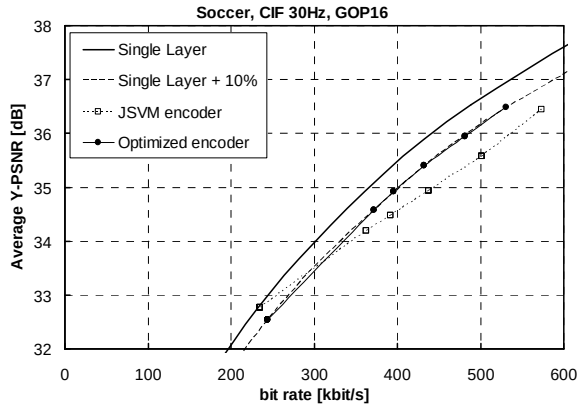


Fig. 11. Experimental results for quality scalable coding of the sequence "Soccer" (CIF resolution, 30 Hz) using an optimized encoder control.

Fig. 11 demonstrates how the coding efficiency of quality scalable coding can be improved by employing the optimized encoder control mentioned in sec. V.B.5. For this simulation, hierarchical B pictures with a GOP size of 16 pictures were used. Quality scalability is achieved by MGS coding without using key pictures. The depicted rate points have been obtained by successively discarding the largest temporal levels of the MGS enhancement layer. It can be seen that coding efficiency can be significantly improved at the high-rate end by tolerating a coding efficiency loss for the lowest rate point. With the optimized encoder control it was possible to limit the bit rate increase compared to single-layer coding at the same fidelity to about 10% over the entire supported bit rate range.

5) SVC-to-H.264/AVC rewriting

The SVC design also supports the creation of quality scalable bit streams that can be converted into bit streams that conform to one of the non-scalable H.264/AVC profiles by using a low-complexity rewriting process [47]. For this mode of quality scalability, the same syntax as for CGS or MGS is used, but two aspects of the decoding process are modified:

1. For the inter-layer intra prediction, the prediction signal is not formed by the upsampled intra signal of the reference layer, but instead the spatial intra prediction modes are inferred from the co-located reference layer blocks, and a spatial intra prediction as in single-layer H.264/AVC coding is performed in the target layer, i.e., the highest quality refinement layer that is decoded for a picture. Additionally, the residual signal is predicted as for motion-compensated macroblock types.
2. The residual prediction for inter-coded macroblocks and for inter-layer intra-coded macroblocks (*base mode flag* is equal to 1 and the co-located reference layer blocks are intra-coded) is performed in the transform coefficient level domain, i.e., not the scaled transform coefficients, but the quantization levels for transform coefficients are scaled and accumulated.

These two modifications ensure that such a quality scalable bit stream can be converted into a non-scalable H.264/AVC bit stream that yields exactly the same decoding result as the quality scalable SVC bit stream. The conversion can be achieved by a rewriting process which is significantly less complex than transcoding the SVC bit stream. The usage of the modified decoding process in terms of inter-layer prediction is signaled by a flag in the slice header of the enhancement layer slices.

VI. SVC High-Level Design

In the SVC extension of H.264/AVC, the basic concepts for temporal, spatial, and quality scalability as described in sec. V are combined. In order to enable simple bit stream adaptation, SVC additionally provides means by which the sub-streams that are contained in a complete scalable bit stream can be easily identified. An SVC bit stream does not need to provide all types of scalability. Since the support of quality and spatial scalability usually comes along with a loss in coding efficiency relative to single-layer coding, the trade-off between coding efficiency and the provided degree of scalability can be adjusted according to the needs of an application. For a further comparison of spatial and quality scalability with single-layer coding the reader is referred to [48].

A. Combined scalability

The general concept for combining spatial, quality, and temporal scalability is illustrated in Fig. 12, which shows an example encoder structure with two spatial layers. The SVC coding structure is organized in dependency layers. A dependency layer usually represents a specific spatial resolution. In an extreme case it is also possible that the spatial resolution for two dependency layers is identical, in which case the different layers provide coarse-grain scalability (CGS) in terms of quality. Dependency layers are identified by a dependency identifier *D*. The spatial resolution must not decrease from one layer to the next. For each dependency layer, the basic concepts of motion-compensated prediction and intra prediction are employed as in single-layer coding; the redundancy between dependency layers is exploited by additional inter-layer prediction concepts as explained in sec. V.B.1.

Quality refinement layers inside each dependency layer are

identified by a quality identifier *Q*. However, when a reference layer for a spatial enhancement layer (dependency layer) contains different quality representations, it needs to be signaled which of these is employed for inter-layer prediction. Therefore, SVC slices include a syntax element, which not only signals whether inter-layer prediction is employed, but also the dependency identifier *D* and the quality identifier *Q* of the corresponding reference layer. For quality refinement layers with a quality identifier $Q > 0$, always the preceding quality layer with a quality identifier $Q - 1$ is employed for inter-layer prediction. In order to limit the memory requirement for storing intermediate representations, all slices of a dependency layer at a specific time instant have to use the same base representation identified by *D* and *Q* for inter-layer prediction.

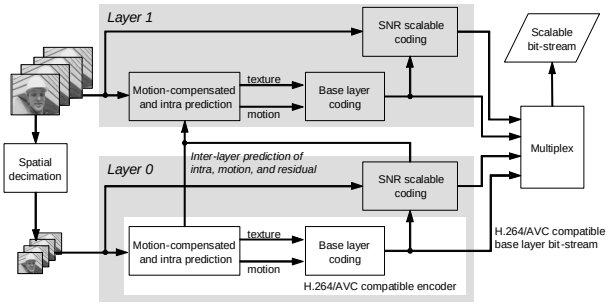


Fig. 12. SVC encoder structure example.

One important difference between the concept of dependency layers and quality refinements is that switching between different dependency layers is only envisaged at defined switching point. However, switching between different quality refinement layers is virtually possible in any access unit. Quality refinements can either be transmitted as new dependency layers (different *D*) or as additional quality refinement layers (different *Q*) inside a dependency layer. This does not change the basic decoding process. Only the high-level signaling and the error-detection capabilities are influenced. When quality refinements are coded inside a dependency layer (identical *D*, different *Q*), the decoder cannot detect whether a quality refinement packet is missing or has been intentionally discarded. This configuration is mainly suitable in connection with hierarchical prediction structures and the usage of key pictures in order to enable efficient packet-based quality scalable coding.

In SVC, all slice data NAL units for a time instant together with zero or more non-VLC NAL units form an access unit. Since inter-layer prediction can only take place from a lower to a higher layer inside an access unit, spatial and quality scalability can be easily combined with temporal scalability. To all slices of an access unit the same temporal level *T* is assigned.

In addition to the main scalability types, temporal, spatial, and quality scalability, SVC additionally supports region-of-interest (ROI) scalability. ROI scalability can be realized via the concepts of slice groups (cp. IV.B), but the shape of the ROI is restricted to patterns that can be represented as a collection of macroblocks.

B. System interface

An important goal for scalable video coding standard is to support easy bit stream manipulation. In order to extract a sub-stream with a reduced spatio-temporal resolution and/or bit rate, all NAL units that are not required for decoding the target resolution and/or bit rate should be removed from a bit stream. For this purpose, parameters like the dependency identifier *D*, the quality identifier *Q*, and the temporal identifier *T* need to be known for each coded slice NAL unit. Furthermore, it needs to be known what NAL units are required for inter-layer prediction of higher layers.

In order to assist easy bit stream manipulations, the 1-byte header of H.264/AVC is extended by additional 3 bytes for SVC NAL unit types. This extended header includes the identifiers *D*, *Q*, and *T* as well as additional information assisting bit stream adaptations. One of the additional syntax elements is a priority identifier *P*, which signals the importance of a NAL unit. It can be used either for simple bit stream adaptations with a single comparison per NAL unit or for rate-distortion optimized bit stream extraction using priority layer information (cp. sec. V.C.3).

Each SVC bit stream includes a sub-stream, which is compliant to a non-scalable profile of H.264/AVC. Standard H.264/AVC NAL units (non-SVC NAL units) do not include the extended SVC NAL unit header. However, these data are not only useful for bit stream adaptations, but some of them are also required for the SVC decoding process. In order to attach this SVC related information to non-SVC NAL units, so-called prefix NAL units are introduced. These NAL units directly precede all non-SVC VCL NAL units in an SVC bit stream and contain the SVC NAL unit header extension.

SVC also specifies additional SEI messages (SEI – Supplemental Enhancement Information), which for example contain information like spatial resolution or bit rate of the layers that are included in an SVC bit stream and which can further assist the bit stream adaptation process. More detailed information on the system interface of SVC is provided in [49]. Information on the RTP payload format for SVC and the SVC file format are given in [50] and [51], respectively.

C. Bit stream switching

As mentioned above, switching between different quality refinement layers inside a dependency layer is possible in each access unit. However, switching between different dependency layers is only possible at IDR access units. In the SVC context, the classification of an access unit as IDR access unit generally depends on the target layer. An IDR access unit for a dependency layer *D* signals that the reconstruction of layer *D* for the current and all following access units is independent of all previously transmitted access units. Thus, it is always possible to switch to the dependency layer (or to start the decoding of the dependency layer) for which the current access unit represents an IDR access unit. But it is not required that the decoding of any other dependency layer can be started at that point. IDR access units only provide random access points for a specific dependency layer. For instance, when an access unit represents an IDR access unit for an enhancement layer and thus no motion-compensated prediction can be used, it is still

possible to employ motion-compensated prediction in the lower layers in order to improve their coding efficiency.

Although SVC specifies switching between different dependency layers only for well-defined points, a decoder can be implemented in a way that at least down-switching is possible in virtually any access unit. One way is to do multiple-loop decoding. That means, when decoding an enhancement layer, the pictures of the reference layers are reconstructed and stored in additional decoded picture buffers although they are not required for decoding the enhancement layer picture. But, when the transmission switches to any of the subordinate layers in an arbitrary access unit, the decoding of this layer can be continued since an additional DPB has been operated as if the corresponding layer would have been decoded for all previous access units. Such a decoder implementation requires additional processing power. For up-switching, the decoder usually has to wait for the next IDR access unit. However, similar to random access in single-layer coding, a decoder can also immediately start the decoding of all arriving NAL units by employing suitable error concealment techniques and deferring the output of enhancement layer pictures (i.e., continuing the output of lower layer reconstructions) until the reconstruction quality for the enhancement layer has stabilized (gradual decoder refresh).

D. Profiles

Profiles and levels specify conformance points to facilitate interoperability between applications that have similar functional requirements. A profile defines a set of coding tools that can be used in generating a bit stream, whereas a level specifies constraints on certain key parameters of the bit stream. All decoders conforming to a specific profile must support all included coding tools.

The SVC Amendment of H.264/AVC specifies three profiles for scalable video coding [10]: Scalable Baseline, Scalable High, and Scalable High Intra. The Scalable Baseline profile is mainly targeted for conversational and surveillance applications that require a low decoding complexity. In this profile, the support for spatial scalable coding is restricted to resolution ratios of 1.5 and 2 between successive spatial layers in both horizontal and vertical direction and to macroblock-aligned cropping. Furthermore, the coding tools for interlaced sources are not included in this profile. For the Scalable High profile, which was designed for broadcast, streaming, and storage applications, these restrictions are removed and spatial scalable coding with arbitrary resolution ratios and cropping parameters is supported. Quality and temporal scalable coding are supported without any restriction in both the Scalable Baseline and the Scalable High profile. Bit streams conforming to the Scalable Baseline and Scalable High profile contain a base layer bit stream that conforms to the restricted Baseline profile and the High profile of H.264/AVC [6], respectively. It should be noted that the Scalable Baseline profile supports B slices, weighted prediction, the CABAC entropy coding, and the 8×8 luma transform in enhancement layers (CABAC and the 8×8 transform are only supported for certain levels), although the base layer has to conform to the restricted Baseline profile, which does not support these tools.

Bit streams conforming to the Scalable High Intra profile, which was mainly designed for professional applications, contain only IDR pictures (for all layers). Beside that, the same set of coding tools as for the Scalable High profile is supported.

VII. Conclusion

In comparison to the scalable profiles of prior video coding standards, the H.264/AVC extension for scalable video coding (SVC) provides various tools for reducing the loss in coding efficiency relative to single-layer coding. The most important differences are:

- The possibility to employ hierarchical prediction structures for providing temporal scalability with several layers while improving the coding efficiency and increasing the effectiveness of quality and spatial scalable coding.
- New methods for inter-layer prediction of motion and residual improving the coding efficiency of spatial scalable and quality scalable coding.
- The concept of key pictures for efficiently controlling the drift for packet-based quality scalable coding with hierarchical prediction structures.
- Single motion compensation loop decoding for spatial and quality scalable coding providing a decoder complexity close to that of single-layer coding.
- The support of a modified decoding process that allows a lossless and low-complexity rewriting of a quality scalable bit stream into a bit stream that conforms to a non-scalable H.264/AVC profile.

These new features provide SVC with a competitive rate-distortion performance while only requiring a single motion compensation loop at the decoder side. Our experiments further illustrate that:

- Temporal scalability: can be typically achieved without losses in rate-distortion performance.
- Spatial scalability: when applying an optimized SVC encoder control, the bit rate increase relative to non-scalable H.264/AVC coding at the same fidelity can be as low as 10% for dyadic spatial scalability. It should be noted that the results typically become worse as spatial resolution of both layers decreases and results improve as spatial resolution increases.
- SNR scalability: when applying an optimized encoder control, the bit rate increase relative to non-scalable H.264/AVC coding at the same fidelity can be as low as 10% for all supported rate points when spanning a bit rate range with a factor of 2-3 between the lowest and highest supported rate point.

Acknowledgment

The authors thank the experts of the Joint Video Team of ISO/IEC MPEG and ITU-T VCEG for their contributions and fruitful discussions. In particular, the authors would like to thank Gary J. Sullivan, Jens-Rainer Ohm, and Jerome Vieron.

APPENDIX
TEST SEQUENCES

The test sequences that are used for simulations in this paper are summarized in TABLE I and TABLE II. All sequences are in YUV 4:2:0 color format, in which the two chroma components are downsampled by a factor of two in each spatial direction. The tables specify the maximum spatial and temporal resolution of the sequences. Sequences with a lower temporal resolution are obtained by frame skipping, and sequences with a lower spatial resolution are obtained by downsampling as specified in the JSVM [22].

TABLE I
HIGH-DELAY TEST SET.

sequence name	abbre-viation	maximum resolution	maximum frame rate	number of pictures
Bus	BU	352x288	30	150
Football	FT	352x288	30	260
Foreman	FM	352x288	30	300
Mobile	MB	352x288	30	300
City	CT	704x576	60	600
Crew	CR	704x576	60	600
Harbour	HB	704x576	60	600
Soccer	SC	704x576	60	600

TABLE II
LOW-DELAY TEST SET.

sequence name	abbre-viation	maximum resolution	maximum frame rate	number of pictures
Group	GR	768x576	30	300
Karsten & Oliver	KO	768x576	30	300
Stefan & Martin	SM	768x576	30	300
Tobias & Cornelius	TC	768x576	30	300
Thomas	TH	768x576	30	300
Uli	UL	768x576	25	250

The test sequences are classified into a high-delay and a low-delay test set. The high-delay test set contains sequences, which have been widely used for testing purposes during the SVC development. The sequences in this set contain different amounts of detail and motion. For low-delay configurations, we used a second, self-recorded test set that is more appropriate for testing low-delay features. Since low-delay is mainly required for interactive video telephone or videoconferencing applications, the low-delay test set consists of a variety of video conferencing sequences.

REFERENCES

[1] ITU-T, "Video codec for audiovisual Services at $p \times 64$ kbit/s," ITU-T Recommendation H.261, Version 1: Nov. 1990, Version 2: Mar. 1993.

[2] ISO/IEC JTC 1, "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 2: Video," ISO/IEC 11 172-2 (MPEG-1 Video), Mar. 1993.

[3] ITU-T and ISO/IEC JTC 1, "Generic coding of moving pictures and associated audio information – Part 2: Video," ITU-T Recommendation H.262 and ISO/IEC 13818-2 (MPEG-2 Video), Nov. 1994.

[4] ITU-T, "Video coding for low bit rate communication," ITU-T Recommendation H.263, Version 1: Nov. 1995, Version 2: Jan. 1998, Version 3: Nov. 2000.

[5] ISO/IEC JTC 1, "Coding of audio-visual objects – Part 2: Visual," ISO/IEC 14492-2 (MPEG-4 Visual), Version 1: Apr. 1999, Version 2: Feb. 2000, Version 3: May 2004.

[6] ITU-T and ISO/IEC JTC 1, "Advanced video coding for generic audiovisual services," ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), Version 1: May 2003, Version 2: May 2004, Version 3: Mar. 2005, Version 4: Sep. 2005, Version 5 and Version 6: June 2006, Version 7: Apr. 2007, Version 8 (including SVC extension): Consented in July 2007.

[7] ITU-T and ISO/IEC JTC 1, "Generic coding of moving pictures and associated audio information – Part 1: Systems," ITU-T Recommendation H.222.0 and ISO/IEC 13818-1 (MPEG-2 Systems), Nov. 1994.

[8] ITU-T, "Narrow-band visual telephone systems and terminal equipment," ITU-T Recommendation H.320, Mar. 1993.

[9] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A transport protocol for real-time applications," RFC 1889, Jan. 1996.

[10] T. Wiegand, G. J. Sullivan, J. Reichel, H. Schwarz, and M. Wien, eds., "Joint Draft 11 of SVC Amendment," Joint Video Team, doc. JVT-X201, Geneva, Switzerland, July 2007.

[11] A. Eleftheriadis, O. Shapiro, and T. Wiegand, "Video conferencing using SVC," *IEEE Transactions on Circuits and Systems for Video Technology*, this issue.

[12] T. Schierl and T. Wiegand, "Mobile video transmission using SVC," *IEEE Transactions on Circuits and Systems for Video Technology*, this issue.

[13] R. Cazoulat, A. Graffunder, A. Hutter, and M. Wien, "Real-time system for adaptive video streaming based on SVC," *IEEE Transactions on Circuits and Systems for Video Technology*, this issue.

[14] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 688-703, July 2003.

[15] N. S. Jayant and P. Noll, "Digital coding of waveforms," Prentice-Hall, Englewood Cliffs, NJ, USA, Mar. 1984.

[16] S.-J. Choi and J. W. Woods, "Motion-compensated 3-d subband coding of video," IEEE Transactions on Image Processing, vol. 8, no. 2, pp. 155-167, Feb. 1999.

[17] B. Pesquet-Popescu and V. Bottreau, "Three-dimensional lifting schemes for motion-compensated video compression," *Proceedings of ICASSP'01*, pp. 1793-1796, Salt Lake City, UT, USA, May 2001.

[18] L. Luo, J. Li, S. Li, Z. Zhuang, and Y.-Q. Zhang, "Motion compensated lifting wavelet and its application in video coding," *Proceedings of ICME'01*, pp. 365-368, Tokyo, Japan, Aug. 2001.

[19] A. Secker and D. Taubman, "Motion-compensated highly scalable video compression using an adaptive 3d wavelet transform based on lifting," *Proceedings of ICIP'01*, vol. 2, pp. 1029-1032, Thessaloniki, Greece, Oct. 2001.

[20] MPEG video sub-group chair, "Registered responses to the call for proposals on scalable video coding," ISO/IEC JTC 1/SC29/WG11, doc. M10569, Munich, Germany, Mar. 2004.

[21] 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, Palma de Mallorca, Spain, Oct. 2004.

[22] J. Reichel, M. Wien, and H. Schwarz, eds., "Scalable Video Model 3.0," ISO/IEC JTC 1/SC 29/WG 11, doc. N6716, Palma de Mallorca, Spain, Oct. 2004.

[23] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560-576, July 2003.

[24] G. J. Sullivan and T. Wiegand, "Video compression – from concepts to the H.264/AVC standard," *Proceedings of IEEE*, vol. 93, no. 1, pp. 18-31, Jan. 2005.

[25] D. Marpe, T. Wiegand, and G. J. Sullivan, "The H.264 / MPEG4 Advanced Video Coding standard and its applications", *IEEE Communications Magazine*, vol. 44, no. 8, pp. 134-144, Aug. 2006.

[26] G. J. Sullivan, H. Yu, S. Sekiguchi, H. Sun, T. Wedi, S. Wittmann, Y.-L. Lee, A. Segall, and T. Suzuki, "New standardized extensions of MPEG4-AVC/H.264 for professional-quality video applications", *Proceedings of ICIP'07*, San Antonio, TX, USA, Sep. 2007.

[27] T. Wiegand, X. Zhang, and B. Girod, "Long-term memory motion-compensated prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 70-84, Feb. 1999.
[28] H. Schwarz, D. Marpe, and T. Wiegand, "Hierarchical B pictures," Joint Video Team, doc. JVT-P014, Poznan, Poland, July 2005.
[29] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MCTF," *Proceedings of ICME'06*, Toronto, Canada, July 2006.
[30] K. Ramchandran, A. Ortega, M. Vetterli, "Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders," *IEEE Transactions on Image Processing*, vol. 13, no. 5, Sep. 1994.
[31] J. Reichel, H. Schwarz, M. Wien, eds., "Joint scalable video model 11 (JSVM 11)," Joint Video Team, doc. JVT-X202, Geneva, Switzerland, July 2007.
[32] M. Flierl, T. Wiegand, B. Girod, "A locally optimal design algorithm for block-based multi-hypothesis motion-compensated prediction," *Proceedings of Data Compression Conference*, Apr. 1998.
[33] H. Schwarz, D. Marpe, and T. Wiegand, "SVC core experiment 2.1: Inter-layer prediction of motion and residual data," ISO/IEC JTC 1/SC 29/WG 11, doc. M11043, Redmond, WA, USA, July 2004.
[34] A. Segall and G. J. Sullivan, "Spatial scalability," *IEEE Transactions on Circuits and Systems for Video Technology*, this issue.
[35] E. François and J. Vieron, "Interlaced coding in SVC," *IEEE Transactions on Circuits and Systems for Video Technology*, this issue.
[36] H. Schwarz, D. Marpe, and T. Wiegand, "Further results for the HHI proposal on combined scalability," ISO/IEC JTC 1/SC 29/WG 11, doc. M11399, Palma de Mallorca, Spain, Oct. 2004.
[37] H. Schwarz, D. Marpe, and T. Wiegand, "Constrained inter-layer prediction for single-loop decoding in spatial scalability," *Proc. of ICIP'05*, Genoa, Italy, Sep. 2005.
[38] H. Schwarz, D. Marpe, and T. Wiegand, "Further results on constrained inter-layer prediction", Joint Video Team, doc. JVT-O074, Busan, Korea, April 2005.
[39] H. Schwarz, D. Marpe, and T. Wiegand, "Independent parsing of spatial and CGS layers," Joint Video Team, doc. JVT-S069, Geneva, Switzerland, March 2006.
[40] H. Schwarz and T. Wiegand, "R-d optimized multi-layer encoder control for SVC," *Proceedings of ICIP'07*, San Antonio, TX, USA, Sep. 2007.
[41] H. Kirchhoffer, H. Schwarz, and T. Wiegand, "CE1: Simplified FGS," Joint Video Team, doc. JVT-W090, San Jose, CA, USA, Apr. 2007.
[42] J.-R. Ohm, "Advances in scalable video coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 42-56, Jan. 2005.
[43] M. Winken, H. Schwarz, D. Marpe, and T. Wiegand, "Adaptive refinement of motion information for fine-granular SNR scalable video coding," *Proceedings of EuMob'06*, Alghero, Italy, Sep. 2006.
[44] H. Schwarz, D. Marpe, and T. Wiegand, "Comparison of MCTF and closed-loop hierarchical B pictures," Joint Video Team, doc. JVT-P059, Poznan, Poland, July 2005.
[45] X. Wang, Y. Bao, M. Karczewicz, and J. Ridge, "Implementation of closed-loop coding in JSVM," Joint Video Team, doc. JVT-P057, Poznan, Poland, July 2005.
[46] I. Amonou, N. Cammas, S. Kervadec, and S. Pateux, "Optimized rate-distortion extraction with quality layers," *IEEE Transactions on Circuits and Systems for Video Technology*, this issue.
[47] A. Segall, "CE 8: SVC-to-AVC bit-stream rewriting for coarse grain scalability," Joint Video Team, doc. JVT-V035, Marrakech, Morocco, Jan. 2007.
[48] M. Wien, H. Schwarz, and T. Oelbaum, "Performance analysis of SVC," *IEEE Transactions on Circuits and Systems for Video Technology*, this issue.
[49] S. Pateux, Y.-K. Wang, M. Hannuksela, and A. Eleftheriadis, "System and transport interface of the emerging SVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, this issue.
[50] S. Wenger and T. Schierl, "RTP payload for SVC," *IEEE Transactions on Circuits and Systems for Video Technology*, this issue.
[51] D. Singer, T. Rathgen, and P. Amon, "File format for SVC," *IEEE Transactions on Circuits and Systems for Video Technology*, this issue.

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.

Detlev Marpe (M'00) received the Dr.-Ing. degree from the University of Rostock, Germany, in 2005 and the Dipl.-Math. degree (with highest honors) from the Technical University Berlin (TUB), Germany, in 1990.

From 1991 to 1993, he was a Research and Teaching Assistant in the Department of Mathematics at TUB. Since 1994, he has been involved in several industrial and research projects in the area of still image coding, image processing, video coding, and video streaming. In 1999, he joined the Fraunhofer Institute for Telecommunications – Heinrich-Hertz-Institute (HHI), Berlin, Germany, where as a Project Manager in the Image Processing Department, he is currently responsible for the development of advanced video coding and video transmission technologies.

Since 1997, he has been an active contributor to the standardization activities of ITU-T VCEG, ISO/IEC JPEG and ISO/IEC MPEG for still image and video coding. From 2001 to 2003, as an Ad-hoc Group Chairman in the Joint Video Team of ITU-T VCEG and ISO/IEC MPEG, he was responsible for the development of the CABAC entropy coding scheme within the H.264/AVC standardization project. He also served as the Co-Editor of the H.264/AVC Fidelity Range Extensions.

In 2004, he received the Fraunhofer Prize for outstanding scientific achievements in solving application related problems and the ITG Award of the German Society for Information Technology. As a co-founder of the Berlin-based daViKo Gesellschaft für audiovisuelle Kommunikation mbH, he was recipient of the Prime Prize of the 2001 Multimedia Start-Up Competition of the German Federal Ministry of Economics and Technology.

Dr. Marpe has published numerous papers on a variety of topics in the area of image and video processing and holds various international patents in this field. His research interests include still image and video coding, image and video communication as well as computer vision, and information theory.

Thomas Wiegand (M'05) is the head of the Image Communication Group in the Image Processing Department of the Fraunhofer Institute for Telecommunications - Heinrich Hertz Institute Berlin, Germany. He received the Dipl.-Ing. degree in Electrical Engineering from the Technical University of Hamburg-Harburg, Germany, in 1995 and the Dr.-Ing. degree from the University of Erlangen-Nuremberg, Germany, in 2000. His research interest include video processing and coding, multimedia transmission, semantic image representation, as well as computer vision and graphics.

From 1993 to 1994, he was a Visiting Researcher at Kobe University, Japan. In 1995, he was a Visiting Scholar at the University of California at Santa Barbara, USA. From 1997 to 1998, he was a Visiting Researcher at Stanford University, USA and served as a consultant to 8x8, Inc., Santa Clara, CA, USA. He is currently a member of the technical advisory boards of the two start-up companies Layered Media, Inc., Rochelle Park, NJ, USA and Stream Processors, Inc., Sunnyvale, CA, USA.

Since 1995, he is an active participant in standardization for multimedia with successful submissions to ITU-T VCEG, ISO/IEC MPEG, 3GPP, DVB, and IETF. In October 2000, he was appointed as the Associated Rapporteur of ITU-T VCEG. In December 2001, he was appointed as the Associated Rapporteur / Co-Chair of the JVT. In February 2002, he was appointed as the Editor of the H.264/AVC video coding standard and its extensions (FREext and SVC). In January 2005, he was appointed as Associated Chair of MPEG Video.

In 1998, he received the SPIE VCIP Best Student Paper Award. In 2004, he received the Fraunhofer Award for outstanding scientific achievements in solving application related problems and the ITG Award of the German Society for Information Technology. Since January 2006, he is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology.