



Scalable Video Coding in a Nutshell

■ **The scalable extension of H.264/AVC (SVC) is the state-of-the-art** scalable video codec jointly developed by the Joint Video Team (JVT) of ITU-T Video Coding Experts Group (VCEG) and ISO/IEC Moving Picture Experts Group (MPEG). Scalable video coding has been an active research and standardisation area for more than 20 years. The functionality of scalability is very attractive due to its capability of reconstructing lower quality or lower resolution videos from partial bit-streams that adapt to various needs of end users as well as varying network conditions. Clearly, different users with different preferences and capabilities can get different quality/resolution videos from a single bit-stream, which greatly promotes resource sharing resulting in better resource utilisation. Several early standards, like MPEG-2 video, H.263, and MPEG-4 visual already include tools to provide different modalities of scalability. However, the scalable profiles of these standards are seldom used. This is because the scalability comes with significant loss in coding efficiency and the Internet was at its early stage. The SVC, which was newly published in July 2007, has several new coding techniques developed and it reduces the gap of coding efficiency with state-of-the-art non-scalable codec while keeping a reasonable complexity increase.

This article first gives an overview of earlier standards and techniques. The basic concept of SVC and the basic tools for providing temporal, spatial and fidelity scalability in SVC are described in detail later. With the understanding of the advances in SVC, and the trend of rapid developments and improvements in network infrastructures, discussion for the future of SVC will be shown at the end of the paper. ■■

Wei Yao
Research Officer, Institute for Infocomm Research

Zheng Guo Li
Research Scientist, Institute for Infocomm Research

Susanto Rahardja
Programme Director, Institute for Infocomm Research

1 INTRODUCTION

With the significant progress in video coding technologies [1]-[7] together with the rapid developments of network infrastructures as well as the exponential growth in storage capacity and computing power, an increasing number of video applications employed a variety of transmission and storage systems that have been widely used in our daily life. Among these applications, video signals could be transmitted over wired/wireless channels with variable bandwidth;

they might be stored on media with different capacities, including a range from low-capacity memory sticks to high-capacity DVDs; they also could be displayed on a variety of devices, ranging from mobile phones with small screens to high-end systems with high-definition displays. In traditional video systems, it is always assumed that the bandwidth required by a video client will be guaranteed. An encoder just needs to compress the input video signal at a bit rate that is less than and close to the predefined bit rate, and the decoder reconstructs the video using all the bits received from the channel. However, in modern video transmission over the Internet, it is almost impossible for the encoder to know the available bandwidth in advance. The video should be encoded over a bit rate range instead of a given bit rate. The conventional non-scalable video coding cannot be used for this type of applications and this gives rise to the need to have a scalable video coding technology.

Scalable video coding involves generating a coded representation (bit-stream) that allows decoding of appropriate subsets to reconstruct complete pictures of resolution or quality commensurate with the proportion of the bit-stream decoded. The minimum bit-stream subset that can be decoded is called base layer. The remaining bits in the bit-stream are called enhancement layer(s) and by decoding the enhancement layer(s) more details are obtained to get the video at higher resolution or quality as compared to base layer. The research on scalable video coding has been an active area for about 20 years. Many early standards, e.g. MPEG-2 Video/H.262 [3] and MPEG-4 Visual [5], have included tools to provide several important scalabilities. However, the scalable profiles of these standards have rarely been used. One reason is due to the characteristics of traditional video transmission systems in which scalabilities is not really necessary. Another main cause for the situation is the fact that scalability always comes along with a significant loss in coding efficiency as well as a large increase in decoder complexity compared to the corresponding non-scalable profiles. In July 2007, a scalable extension of H.264/MPEG-4 AVC (Advanced Video Coding) was jointly published by MPEG and ITU-T Video Coding Experts Group (VCEG), which makes the scalable extension to be the state-of-the-art scalable video codec. Several new coding techniques were developed in the scalable extension and the gap of coding efficiency has been reduced with state-of-the-art non-scalable codec while the complexity increase is reasonably maintained.

This article will first give an introduction of essential scalability types as well as an overview of scalability techniques in earlier standards. The basic concept in scalable extension of H.264/MPEG-4 AVC and the basic tools for providing temporal, spatial and fidelity scalability in the scalable extension will also be described later. With the understanding of scalabilities, advances in state-of-the-art scalable extension of H.264/MPEG-4 AVC and reasons why scalable codec always comes with coding efficiency drop as compared with single layer coding will be discussed.

2 SCALABILITY BASICS

2.1 Scalability Approaches

There are two popular approaches to achieve scalability, one is simulcast coding and the other is scalable video coding. The details are given as below.

In Simulcast Coding, two or several bit-streams targeting at different decoded qualities or resolutions are tied together for the purpose of parallel transmission, as depicted in Figure 1. In other words, each layer of video representing a resolution or quality is coded or transcoded independently. Thus any layer can be decoded by a single-layer (non-scalable) decoder and it is assumed that independent decoders would be used to decode each layer. Total available bandwidth is simply partitioned depending on the quality desired for each independent layer that needs to be coded.

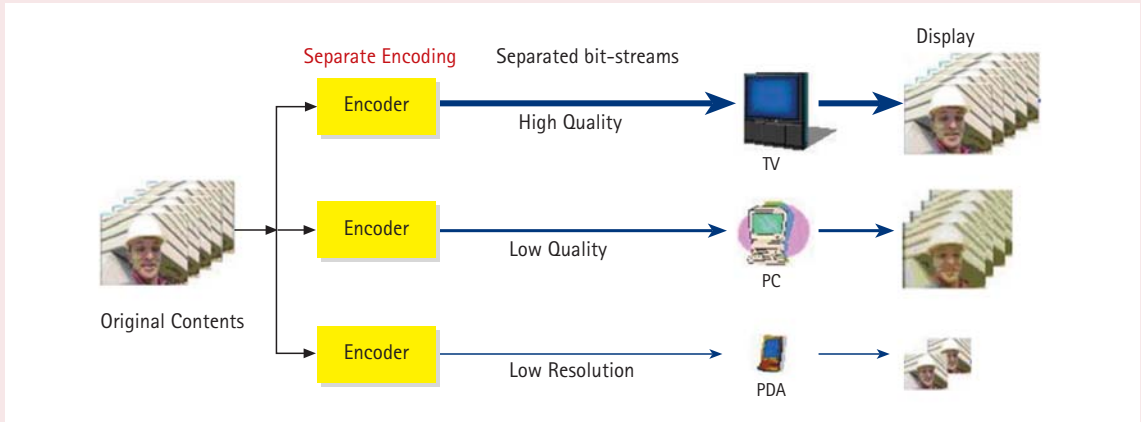


Figure 1: Examples of scalability approach Simulcast Coding

In Scalable Video Coding, only one layer - the base layer is coded independently whereas other layers are coded dependently, each following layer coded with respect to the previous layers. Therefore, only a single bit-stream is generated but parts of it can be extracted in a way that the resulting sub-stream forms another valid bit-stream for a given decoder, as shown in Figure 2. Scalable coding is generally more efficient than simulcast. Except for the independently coded base layer, each enhancement layer is able to reuse some of the bandwidth assigned to the previous layer. The exact amount of increased efficiency is dependent on the specific technique used, the number of layers and the bandwidth partitioning used. But the increase in efficiency comes at the expense of some increase in complexity as compared to simulcast coding. This is the tradeoff in scalable video coding. Different scalabilities offer different tradeoffs, and in general some are more suitable for one set of applications while others are better suited for another set of applications and so forth. Various types of scalability and a brief explanation of their intended applications are provided in next section.

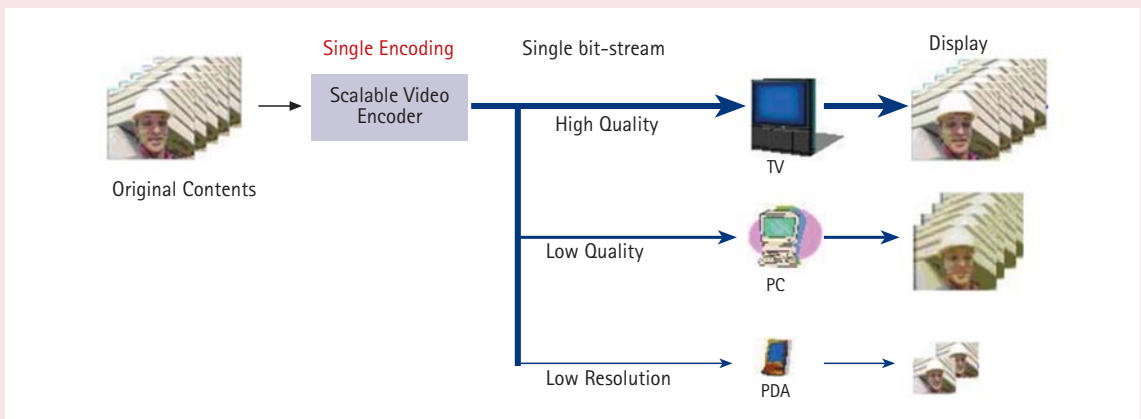


Figure 2: Examples of scalability approach of Scalable Coding

2.2 Scalability Types

Signal-to-Noise Ratio (SNR) Scalability

SNR scalability involves generating two or more layers of the same spatiotemporal resolution but different video qualities from a single video source such that the base layer is coded to provide the basic video quality and the enhancement layer(s) when added back to the base layer reconstruct a higher quality reproduction of the input video. Since the enhancement layer is said to enhance the signal-to-noise ratio of the base layer, this type of scalability is called SNR scalability. This scalability is a tool intended for use in many video applications, for example, telecommunications and multiple quality video services with standard TV and High-Definition TV; Multi-quality video-on-demand services; Error-resilient video over asynchronous transfer mode network (ATM) and other networks where the base layer carries the most significant information and the enhancement layer carries additional data that is less critical; and so on.

Spatial Scalability

Spatial scalability involves generating two or more layers with different spatial resolutions from a single video source such that the base layer is coded by itself to provide the basic spatial resolution and the enhancement layer(s) employ the spatial interpolated base layer and carry higher or full spatial resolution of the video source. Spatial scalability supports interoperability between applications using different video formats, e.g. the base layer can use QCIF resolution (176x144) at 4:2:0 while the enhancement layer can be kept at CIF resolution (356x288) at 4:2:2. It also enables interworking with different standards, between various telecommunication applications, and providing resilience to transmission errors in a similar way as SNR scalability.

Temporal Scalability

Temporal scalability involves partitioning of video frames into layers, in which the base layer is coded to provide the basic frame rate and the enhancement layer(s) is coded with temporal prediction with respect to the base layer. The layers may have either the same or different temporal resolutions, which, when combined, provide full temporal resolution at the decoder. This scalability can be used in a range of diverse video applications from telecommunications to HDTV. For example, systems migration to higher temporal resolutions from that of lower temporal resolution systems may require temporal scalability to provide backward compatibility. In many cases, the lower temporal resolution video systems may be either the existing systems or the less expensive early systems. The more sophisticated systems may then be introduced gradually. Besides these applications, temporal scalability may also be useful in applications such as a software decoding environment where the decoding processor may not be powerful enough to decode video at the full frame rate or may be sharing its processing power among a number of different tasks.

Combined Scalability

Individual scalabilities can be combined to form mixed scalability for certain applications. To name a few, consider for instance, a scenario of a video transmission service with heterogeneous clients, which request same video content but with different resolutions, qualities and frame rates. With a proper configured scheme providing combined scalability, the source content just needs to be encoded once for the highest requested resolution, frame rate and

bit rate, forming a scalable bit-stream from which representations of lower resolution, lower frame rate and lower quality can be obtained by partial decoding. Combined scalability is highly desired for application such as surveillance, in which the video content will not only be viewed on different display systems, ranging from small screen video phones to high definition monitors, but also need to be stored and archived. The high resolution/quality part in the bit-stream could be deleted after sometime and only the low quality copies to be stored for archiving purpose. Another important application, as mentioned in all previously introduced individual scalabilities, is to use combined scalability together with unequal error protection scheme for video transmission in networks with unpredictable throughput variations. By a strong protection of lower layers, the connection will not be completely interrupted in the presence of transmission error and a base quality with graceful degradation can still be received.

3 SCALABILITY IN EXISTING VIDEO CODING STANDARDS

Early video compression international standards such as ITU-T H.261 [1] and ISO/IEC MPEG-1 [2] did not have any scalability mechanisms. This was due to the specific communication applications at that time, e.g. conversational services, did not require any scalable functionalities. ISO/IEC MPEG-2, which is identical to ITU-T H.262 [3], was the first general-purpose video compression standard that includes a number of tools providing scalability. The later video codec of the ISO/IEC MPEG-4 standard [5] provides even more flexible scalability tools within a more generic framework, especially the SNR scalability with fine granularity scalability (FGS) at the level of video objects providing a continuous scalability in which the enhancement bit-stream can be truncated into any number of bits within each frame. In this section, scalable mechanisms in MPEG-2 and MPEG-4 are discussed.

3.1 Scalabilities in MPEG-2 Video

Following the universal success of H.261 and MPEG-1 video codecs, there was a growing need for a video codec to address a wide variety of applications. Considering the similarity between H.261 and MPEG-1, ITU-T and ISO/IEC made a joint effort to devise a generic video codec, which targeted at coding of video for transmission over the Broad-band Integrated Service Digital Networks (B-ISDN) using ATM transport. The devised generic codec was finalised in 1994, and takes the name of MPEG-2/H.262. MPEG-2 was the first standard to include implementations of layered coding where enhancement layer performs differential encoding with reference to the base layer. All dimensions of scalabilities mentioned above are supported in MPEG-2.

SNR scalability in MPEG-2

Two layers in SNR scalability are at the same spatial and temporal resolutions but produce different qualities. The enhancement layer includes DCT coefficient refinement, and when added back base layer, a higher quality video can be regenerated. The video encoder first quantizes the DCT coefficients to a given accuracy with a quantization parameter QP_b ; the quantized coefficients are then coded by variable length coder (VLC) and transmitted as the base layer bit-stream. The quantization error introduced by the first quantizer is quantized again with a finer quantization parameter QP_e (where $QP_b > QP_e$), and then coded by VLC and transmitted as the enhancement layer bit-stream. Side information required by the decoder, such as motion vectors, is transmitted only in the base layer.

To decode the combined base and enhancement layers, both layers must be received. In MPEG-2 SNR scalability encoder, the refined DCT coefficients in enhancement layer are fed back into the motion compensation loop to regenerate the higher quality reconstructed pictures which are to be used as the reference pictures for motion estimation/motion compensation. The prediction structure is shown in Figure 3. If only partial information is available to generate prediction at decoder, e.g. the reconstruction of the video from the base layer only, it can result in a mismatch between the encoder and decoder prediction loops. This causes drift on the reconstructed video and there is a tendency for *drift* to accumulate over P-frames, which results in a drop of more than 1dB in coding efficiency as compared to single layer coding.

Data partitioning is a simpler method to implement SNR scalability. It is just a technique to partition a single-layer coded bit-stream into layers. Therefore, it can be realised that both the base and enhancement layer coefficients are used at the encoding prediction loop at the encoder. Potential accumulation of drift makes the quality of base layer to be significantly lower by as much as 3dB or more [15] than that of SNR scalability, especially at low bit rate. This is because, at lower base layer bit rates, data partitioning can only retain DC and possibly one or two AC coefficients and reconstructed pictures with these few coefficients would be very blocky. However, by combining base and enhancement layers in the case of no errors, the quality obtained by data partitioning could be higher than MPEG-2 SNR scalability due to its low complexity and close to zero overhead as compared with the single-layer encoder.

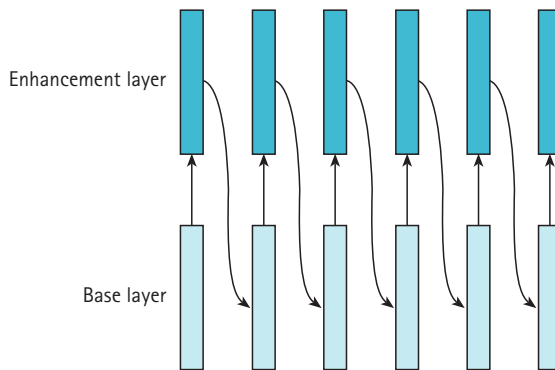


Figure 3: Prediction structure for SNR scalability in MPEG-2

Spatial Scalability in MPEG-2

As different from SNR scalability, spatial scalability in MPEG-2 is characterised by the use of decoded pictures from base layer as a prediction in the enhancement layer. Two coder loops operate at different resolutions are used to produce the base and enhancement layers. The input video is first spatial down-sampled in both horizontal and vertical directions to produce a reduced picture resolution. The base layer produces a bitstream which may be decoded as the non-scalable case. The up-sampled locally decoded picture from the base layer is used as a prediction for the blocks in the enhancement layer. This prediction is in addition to the prediction from the upper level's motion

compensated prediction. The adaptive prediction is implemented by the weighting function w_1 as shown in Figure 4 and this is called spatiotemporal weighted prediction. For each macroblock (MB), a spatially interpolated decoded lower layer signal is adaptively combined with motion compensated prediction using a set of predefined weights from a selected table. Figure 4 illustrates the principle of spatiotemporal weighted prediction, for each 16x16 block in the enhancement layer, the corresponding decoded 8x8 block of the base layer is up-sampled spatially to 16x16 and combined with a 16x16 temporal motion compensated block using a choice of weights to generate 16x16 candidate blocks for weighted spatiotemporal prediction.

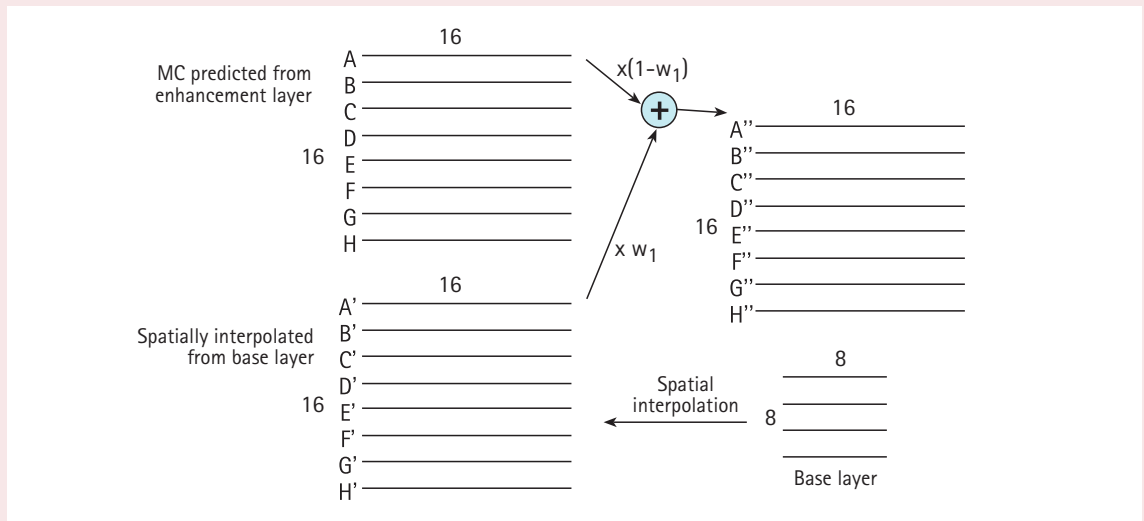


Figure 4: Principle of spatiotemporal weighted prediction in spatial scalability

Temporal Scalability in MPEG-2

Since in temporal scalability the input video frames are simply partitioned between the base and enhancement layer encoders, the encoder need not be more complex than a single-layer encoder. For example, the single layer encoder may be switched between the two base and enhancement modes to generate a base and enhancement bit-stream alternately and the decoder can be reconfigured to decode the two bit-stream alternately too. In fact the B-frames in MPEG-2 provide a very simple temporal scalability that is encoded and decoded alongside the anchor I- and P-frames within a single codec. I- and P- frames can be regarded as the base layer, and the B-frames become the enhancement layer, as the example shown in Figure 5. The arrow shows the direction of motion compensated prediction. Decoding I- and P-frames along will result in the base layer with low temporal resolution, and when added to the decoded B-frames the temporal resolution is enhanced to its full size.

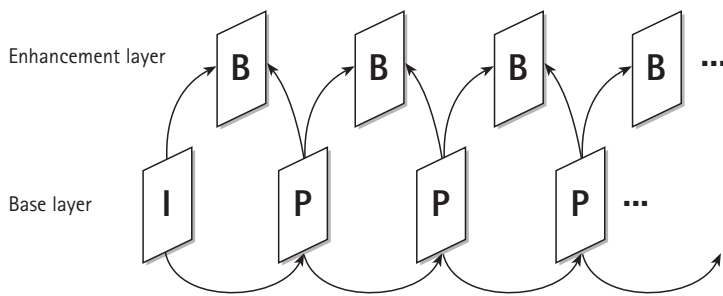


Figure 5: An example of motion compensated prediction from temporal scalability

Hybrid Scalability in MPEG-2

The MPEG-2 standard allows combining two scalabilities at a time from among spatial scalability, the SNR scalability, and the temporal scalability. It is useful in more demanding applications requiring two or more layers.

As a summary, MPEG-2 was the first standard to include implementations of layered coding, where the standalone enhancement layer information without the base layer is useless because differential encoding is performed with reference to the base layer. All dimensions of scalability are supported in MPEG-2. However, the number of scalable bit-stream layers is generally restricted to a maximum of three in any of the existing MPEG-2 profiles.

3.2 Scalabilities in MPEG-4 Visual

Influenced by the well-established video compression standards and the new ideas of content coding, the MPEG-4 video coding activities began in 1995 with the aim of combining the benefits of both approaches: high compression over a broad range of bit rates and innovative content-based functionalities. There are actually two parts on video coding in MPEG-4 standards, MPEG-4 Part 2 Visual [2] and MPEG-4 Part 10 Advanced Video Coding [6] (which is also referred to as H.264). In this section, only MPEG-4 Part 2 Visual is discussed.

In addition to previous standards that addressed only coding efficiency, the MPEG-4 Part 2 visual provides several additional functionalities, such as coding of arbitrarily shaped objects; efficient compression of video sequences and still images over a wide range of bit rates; spatial, temporal, and quality scalability and so on. These advanced features greatly aid content creators in generating multimedia content. In particular, the ability to code objects of arbitrary shape and size goes beyond the scope of all the previous video and image coding standards [5]. And for scalability in MPEG-4 Visual, the scalability tools are more flexible than MPEG-2.

Besides the normal frame level scalability, they can also be at the level of video objects [5]. In object-based scalability case, the base layer is decoded by using non-scalable tools and the enhancement layer contain the complementary information required to reconstruct the higher resolution/quality signals, or the remaining number of video objects. In this section, the specific scalabilities in MPEG-4 Visual will be introduced and only the case of a single enhancement layer is described for simplicity.

SNR Scalability in MPEG-4 Visual

The SNR scalability in MPEG-4 Visual is fundamentally different from MPEG-2 Video.

In MPEG-4 SNR Fine Granularity Scalability (FGS), the reconstruction error of the base layer is encoded in the enhancement layer using a bit plane representation of the DCT coefficients, as shown in Figure 6. The absolute values of the DCT coefficients (Figure 6a) are arranged according to their binary representation (Figure 6b). Then the coefficient information is grouped bit plane by bit plane, starting from the most significant bit plane (MSB) to the least significant bit plane (LSB). This is followed by run-length coding of the bits for each bit plane (Figure 6c). The MSBs of the enhancement layer signal are encoded into the bit-stream for all MBs, followed by the second significant bit planes and so on. Thus it is possible to stop the transmission of the enhancement layer data at any point.

MPEG-4 FGS is defined only for rectangular, non-arbitrarily shaped video objects. To provide functionality similar to the adaptive quantization matrix and to the MB adaptive quantization step size used in the coding of base layer, MPEG-4 FGS applies frequency weighting and selective enhancement shifting factors before encoding the bit planes.

For certain MBs in the Video Object Plane (VOP) (region of interests for which a selectively enhanced representation by the encoder may improve the quality impact) or for the most relevant DCT coefficients, a finer quantization is applied in the FGS enhancement layer. This is done by simply shifting up the corresponding coefficients by one or more bit planes according to a key or mask that is also sent to decoder.

As a result, the most important MBs or DCT coefficients in the enhancement layer are coded first, and thus have a higher probability of being used by decoder if bit-stream is truncated at a certain stage for any reason.

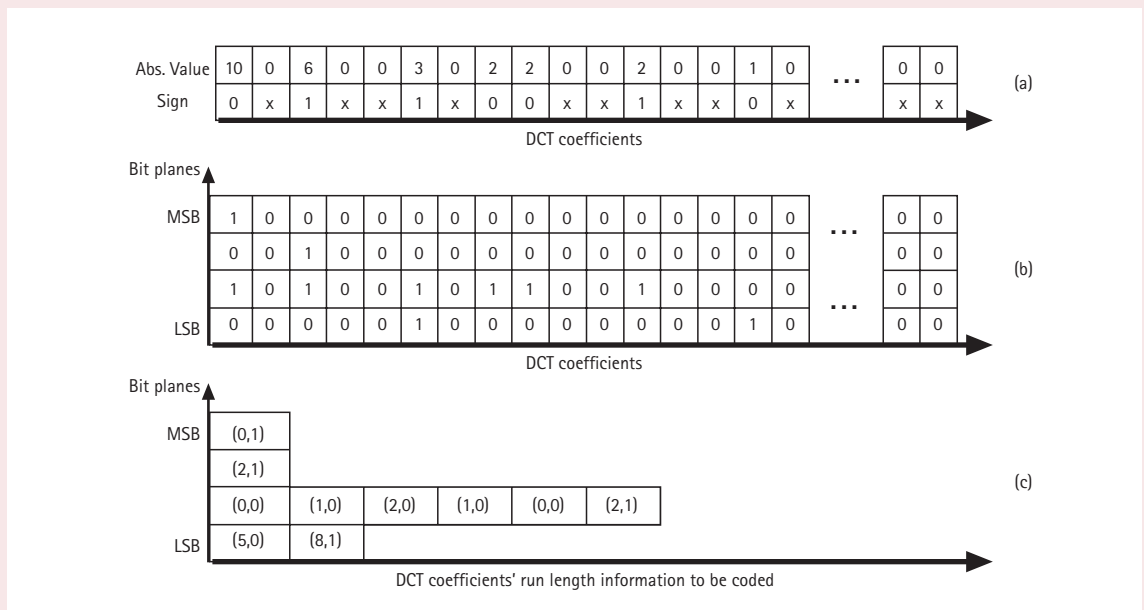


Figure 6: FGS bit plane encoding

Note that the reconstructed enhancement layer frames are used in the motion compensated prediction loop in MPEG-2 encoder, which results in a drift when only the base layer is decoded. In order to avoid the drift of the prediction signals between the encoder and a base layer decoder in MPEG-4 Visual, the reference frames for the motion compensation are usually the reconstructed frames in the base layer. Figure 7 illustrates the prediction structures of MPEG-4 FGS. Comparing Figure 3 and Figure 7, it can be observed that MPEG-2 is designed in a way that the enhancement layer would not be affected by drift problem and thus enhancement layer is optimised while MPEG-4 FGS is designed to optimise the base layer.

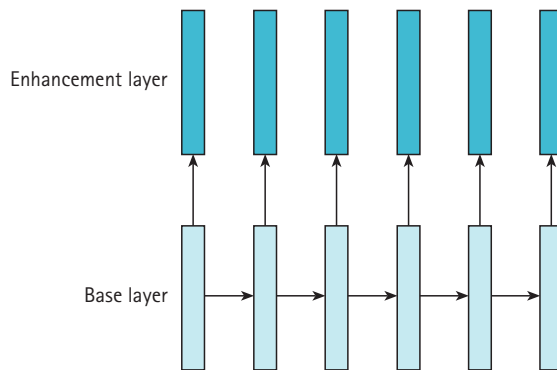


Figure 7: Prediction structure for SNR scalability in MPEG-4 FGS

Spatial Scalability in MPEG-4 Visual

The coding structure for spatial scalability in MPEG-4 is very similar to that in MPEG-2. Unlike MPEG-2, the spatial scalability of MPEG-4 operates on VOPs at different resolutions. The base layer decoder reconstructs low-resolution VOPs using non-scalable coding tools. These VOPs are then up-sampled and used to predict the higher resolution enhancement layer VOPs in combination with the enhancement layer's motion compensated prediction. The combined prediction, which is different from the weighted spatiotemporal prediction (see Section 3.1 Figure 4) in MPEG-2, adopts a bi-directional prediction pattern.

Although MPEG-2 video allows spatial scalability of both interlaced and progressive video, MPEG-4 spatial scalability is only defined for non-interlaced, progressively scanned video. Because MPEG-4 allows the coding of arbitrarily shaped video objects, spatial scalability is also supported for arbitrarily shaped objects. In this case, the binary shape information for the enhancement layer is generated and coded in enhancement layer. For the bi-directional prediction for enhancement layer, the base layer texture must be padded before the bilinear texture up-sampling [5].

Temporal Scalability in MPEG-4 Visual

The prediction of enhancement layer I-, P- and B-VOPs in MPEG-4 temporal scalability is performed in the same way as for the I-, P- and B-frames in MPEG-2 Video. The base layer decoder reconstructs low temporal resolution VOPs,

which are then used together with the previously reconstructed enhancement layer VOPs in the enhancement layer's motion compensated prediction. In MPEG-4 Visual, temporal scalability can be combined with either spatial scalability or FGS but the combination of spatial scalability and FGS is not possible.

As a conclusion, the scalable coding in MPEG-4 Visual is featured by the object-based scalability. Both spatial and temporal scalability can be combined with arbitrary shape coding to provide scalability on arbitrary shaped objects and FGS can be applied on rectangular VOPs. MPEG-4 FGS is the most well known scalability among all the scalabilities in MPEG-4 Visual. In addition to the scalabilities mentioned above, data partitioning is also adopted in MPEG-4 to enable better resynchronisation and error localisation.

4 SCALABLE EXTENSION OF THE H.264/AVC STANDARD

Responding to the demand of scalability in video coding, ISO/IEC 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 [9] represented scalable video codecs based on a 3-D wavelet transform, while the remaining two proposals were extension of H.264/MPEG-4 AVC [6].

After a six month evaluation phase in which several subjective tests for a variety of conditions were carried out, the proposals were carefully analysed with consideration to having a successful future standard, the scalable extension of H.264/MPEG-4 AVC as proposed by HHI were chosen as the starting point of MPEG's Scalable Video Coding (SVC) project in October 2004.

In January 2005, MPEG and ITU-T Video Coding Experts Group (VCEG) agreed to jointly finalise the SVC project as an Amendment of their H.264/MPEG-4 AVC standard, named as scalable extension of H.264/AVC standard. The standardisation activity of this scalable extension was completed and the standard was published in July 2007, which completed the milestone for scalable extension of H.264/AVC to become the state-of-the-art scalable video codec in the world. In this section, scalable extension of H.264/AVC and its coder structure for the supported scalabilities are discussed.

4.1 Requirements for a Successful Scalable Video Codec

Considering the requirements 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 has to be characterized by the following features and requirements [8]:

- Similar coding efficiency compared to single-layer coding;
- Low increase in decoding complexity compared to state-of-the-art single-layer decoding;
- The decoding complexity that scales with the decoded spatio-temporal resolution and bit rate;
- Support of temporal, spatial, and fidelity scalability; and
- Support of simple bit-stream adaptations.

In any case, the coding efficiency should be superior to that of simulcasting the supported spatiotemporal resolutions and bit-rates using a state-of-the-art single-layer video codec. Comparing to single layer coding, bit-rate increases of 10% [8] for the same fidelity might be tolerable depending on the needs of an application and the supported degree of scalability.

4.2 General Coder Structure

The basic design of H.264/MPEG-4 AVC scalable extension is also layered video codec. In general, the coder structure as well as the coding efficiency depends on the scalability space that is required by an application. A typical coder structure with three spatial layers is illustrated in Figure 8.

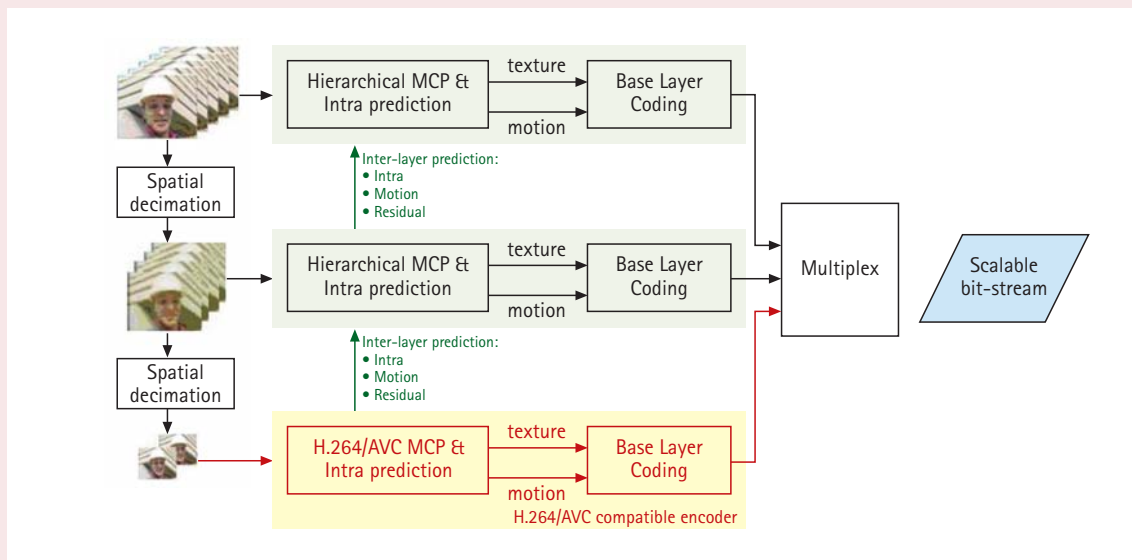


Figure 8: The general coder structure of a three spatial layers scalable extension of H.264/AVC [12]

In each spatial or coarse-grain SNR layer, the basic concepts of motion-compensated prediction and intra prediction are derived from H.264/MPEG4-AVC. The redundancy between different layers is exploited by new concepts of inter-layer prediction that include prediction mechanisms for motion parameters as well as texture data (intra and residual data).

A base representation of the input pictures of each layer is obtained by transform coding similar to that of H.264/MPEG4-AVC, the corresponding network abstraction layer (NAL) units contain motion information and texture data; the NAL units of the lowest layer are compatible with single-layer H.264/MPEG4-AVC [6].

Inter-Layer Motion Prediction

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

Inter-Layer Residual Prediction

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

Inter-Layer Intra Prediction

Furthermore, an additional intra MB mode is introduced, in which the prediction signal is generated by up-sampling the co-located reconstruction signal of the lower layer. For this prediction it is generally required that the lower layer is completely decoded including the computationally complex operations of motion-compensated prediction and de-blocking.

However, this problem can be circumvented when the inter-layer intra prediction is restricted to those parts of the lower layer picture that are intracoded [7]. With this restriction, each supported target layer can be decoded with a single motion compensation loop.

4.5 SNR Scalability in Scalable Extension of H.264/AVC

For SNR scalability, coarse-grain SNR scalability (CGS) and medium-grain SNR scalability (MGS) are distinguished in current scalable extension of H.264/AVC.

Coarse-Grain SNR Scalability

Coarse-grain SNR scalable coding is achieved using the concepts for spatial scalability. The same inter-layer prediction mechanisms are employed; the only difference is that for CGS the up-sampling operations are omitted.

The CGS 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. Furthermore, the CGS concept becomes less efficient when the relative rate difference between successive CGS layers gets smaller [13].

Medium-Grain SNR Scalability

In order to increase the granularity for SNR scalability, scalable extension of H.264/MPEG-4 AVC provides a variation of CGS approach, which uses the quality identifier Q for quality refinements. This method is referred to as MGS and allows the adaptation of bit stream adaptation at an NAL unit basis. With the concept of MGS, any enhancement layer NAL unit can be discarded from a quality scalable bit-stream and thus packet based SNR scalable coding is obtained. However, it requires a good controlling of the associated drift.

As mentioned, the prediction structure of FGS in MPEG-4 Visual was chosen in a way that drift is completely omitted. As illustrated in Section 3.2, motion compensation prediction in MPEG-4 FGS is usually performed using the base layer reconstruction for reference and thus loss of any enhancement packet does not result in any drift on the motion compensated prediction loops between encoder and decoder.

The drawback of this approach, however, is the significant decrease of enhancement layer coding efficiency in comparison to single layer coding, because the temporal redundancies in enhancement layer cannot be properly removed.

For SNR scalability coding in MPEG-2 Video, the other extreme case was specified. The highest enhancement layer reconstruction is used in motion compensated prediction. This ensures a high coding efficiency as well as low complexity for the enhancement layer. However, any loss or modification of a refinement packet results in a drift that can only be stopped by intra frames.

For the MGS in scalable extension of H.264/AVC, an alternative approach, which allows certain amount of drift by adjusting the tradeoff between drift and enhancement layer coding efficiency is used. The approach is designed for SNR scalable coding in connection with hierarchical prediction structures.

For each picture, a flag is transmitted to signal whether the base representations or the enhancement representations are employed for motion compensated prediction. Picture that only uses the base representations ($Q=0$) for prediction is also referred as key pictures [7]. Figure 10 illustrates how the key picture can be combined with hierarchical prediction structures.

All pictures of the coarsest temporal level are transmitted as key pictures, and thus no drift is introduced in the motion compensated loop of temporal level 0. In contrast to that, all temporal refinement pictures are using the highest available quality pictures as reference in motion compensated prediction, which results in high coding efficiency for these pictures. Since key pictures serve as the resynchronisation point between encoder and decoder reconstruction, drift propagation can be efficiently contained inside a group of pictures. The tradeoff between drift and enhancement layer coding efficiency can be adjusted by the choice of GOP size or the number of hierarchy stages.

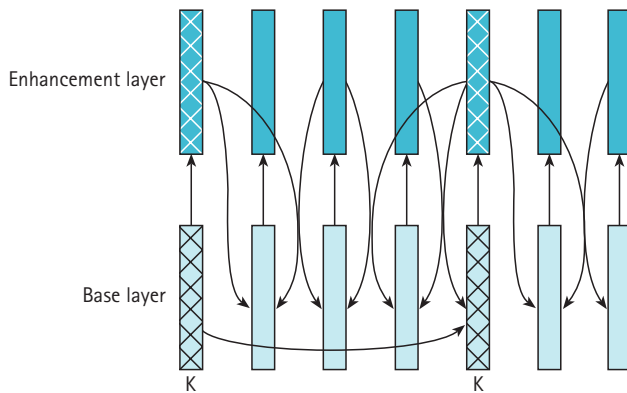


Figure 10: Key picture concept of H.264 scalable extension for hierarchical prediction structure demonstrating the tradeoff between drift and enhancement layer coding efficiency. K denotes the key pictures [12].

4.6 Performance of Scalable Extension of H.264/AVC

The concepts for temporal, spatial, and SNR scalability can be easily combined. In Figure 11 [12], the coding efficiency of combined scalable coding is compared to the coding efficiency of single-layer, purely spatial scalable, and purely SNR scalable coding for the sequence "Soccer". The intra period was generally set to 1.07 s (64 pictures at 60Hz); and for all encodings a dyadic hierarchical prediction structure with a GOP size of 32 pictures at 60Hz is used. Since temporal scalability with resolution of 1.875Hz to 60Hz (4CIF) is supported in the same manner in all bit-streams, it has not been tested separately. The black curves show the coding efficiency of single layer coding; each point represents a separate bit-stream.

For SNR scalability, a separate bit-stream has been generated for each spatial resolution and is represented by the corresponding red curves. Similarly, the blue curves show the coding efficiency of spatial scalable coding. Three bit-streams have been generated, and each of these bit-streams includes the lowest, the middle, or the highest plotted rate point for each spatial resolution. The coding efficiency of combined scalable coding, for which all plotted rate points with different spatiotemporal resolutions are supported in a single bit stream, is represented by the green curves.

The combined scalable bit-stream provides the same amount of extractable sub-streams as the three SNR scalable bit-streams, and it provides even more extractable rate points with different spatiotemporal resolutions than the three spatial scalable bitstreams or all generated single layer bit-streams. It can be seen on the example of Figure 11, the coding efficiency of scalable extension of H.264/AVC bit-streams changes with the range of supported spatiotemporal-rate points. Each additional representation of the video source requires the coding of motion and texture data to be spread over several data partitions, which results in a suboptimal coding efficiency for each supported spatiotemporal rate points in comparison to single layer coding. But due to the flexibility of the SVC design the tradeoff between the amounts of supported spatiotemporal rate points and the coding efficiency can be adjusted according to the needs of an application.

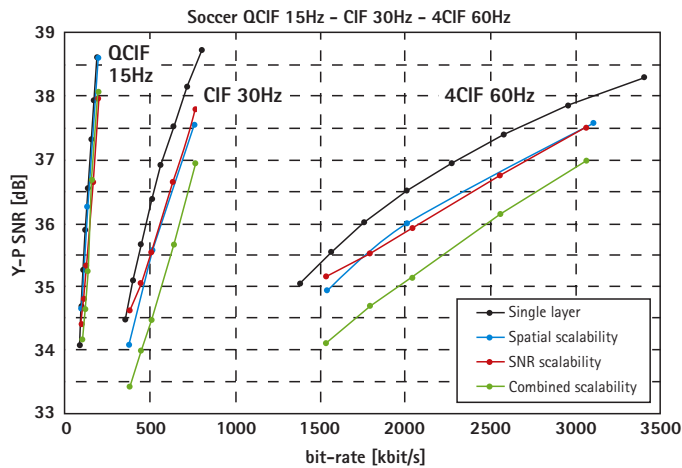


Figure 11: Performance of combined scalability in H.264 scalable extension for sequence "Soccer" [12]

5 DISCUSSION

Major reasons for the performance gap between scalable coder and single layer coder.

In video coding, a lack of coding efficiency in comparison with single layer coding can generally be observed in combining scalable coding with the approach of hybrid motion compensated prediction and block transform coding, as implemented in most of today's standards. The problem mainly comes from the recursive structure of the prediction loop, which causes a drift problem whenever incomplete information is decoded.

Taking the SNR scalability in MPEG-2 for example, if the base layer uses the decoded enhancement layer data in the encoder prediction loop, drift may not take place in the decoded enhancement layer but in the base layer. If using a simpler configuration without feedback of decoded enhancement layer data in the base encoder, drift will also occur in the enhancement layer at the decoder.

The quality gap between MPEG-2 and non-scalable single layer coding is usually more than 1dB. Although MPEG-4 FGS adopts a structure that completely omits drift, the structure results in a huge decrease of enhancement layer coding efficiency to as large as 2dB [14] at the high end of bit rate range. In addition to the drift and drift propagation problems, the layered differential coding structure, in which redundancies between layers cannot be completely removed, is another reason for the low coding efficiency in scalable coder. It is caused by the following possible factors:

- Certain important information, such as headers must be duplicated in both base and enhancement layers;
- Run-lengths, used to index VLC code tables, are in general longer for scalable coders due to the absence of certain non-zero coefficients that are encoded only in the base and others that are encoded only in the enhancement layer; and

- The combined bit allocation for some coefficients appearing in both the base and enhancement layers may exceed that required in an equivalent single layer coder.

Without good drift control schemes and good redundancy removal technologies, early standards have led to a situation that even though numerous scalable tools were available in early video coding standards, there was no wide acceptance in the market of the prospective applications.

Advances in Scalable Extension of H.264/AVC

In comparison to the early scalable standards, scalable extension of H.264/MPEG-4 AVC provides various tools for improving efficiency relative to single-layer coding. The key improvements that make the scalable extension of H.264/MPEG-4 AVC superior than all scalable profiles in early standards are listed below:

- 1) The employed hierarchical prediction structure that provides temporal scalability with several levels improves the coding efficiency and effectiveness of SNR and spatial scalable coding.
- 2) The concept of key pictures with hierarchical prediction structure efficiently controls the tradeoff between drift and enhancement layer coding efficiency. It provides a basis for efficient SNR scalability, which could not be achieved in all previous standards.
- 3) New mechanism for inter-layer prediction of motion and residual information improves coding efficiency of spatial and SNR scalability. In all previous standards, only residual information can be refined at enhancement layers. Several layers may share one motion vector field that is generated to optimise a specific layer, either base layer or enhancement layer but not both. Without motion refinement, if the bit rate scalable range is very large between base layer and enhancement layer, there would not be enough motion information to support refined residual information and a significant drop in the coding efficiency would be observed. In scalable extension of H.264/AVC, both residual and motion information can be refined at enhancement layers [7] [10], and the correlations between layers are well exploited which leads to a high coding efficiency.
- 4) The coder structure is designed in a more flexible way such that any layer can be configured to be the optimisation point. MPEG-2 is designed in the sense that enhancement layer is always optimised but the base layer may suffer from a serious drift problem that causes significant quality drop. MPEG-4 FGS, on the other way round, usually coded in a way to optimised base layer and the coding efficiency of enhancement layer is much lower than single layer coding. In scalable extension of H.264/MPEG-4 AVC, the optimum layer can be set to any layer with a proper configuration [10][11]. This is very useful in applications that customers' interests must be taken into considerations. The video source can be adaptively coded according to customers' interest so that the coding could be optimised to provide full customer satisfaction and maximum organisation's profit.
- 5) Single motion compensated loop decoding provides a decoder complexity close to single layer decoding.

In conclusion, with the advances mentioned above, the state-of-the-art scalable video codec – scalable extension of H.264/MPEG-4 AVC, has enabled profound performance improvements for both scalable and single layer coding. Results of the rate-distortion comparison show that scalable extension of H.264/MPEG-4 AVC clearly outperforms early video coding standards, such as MPEG-4 ASP [13]. Although scalable extension of H.264/MPEG-4 AVC still comes at some costs in terms of bit rate or quality, the gap between the state-of-the-art single layer coding and scalable extension of H.264/MPEG-4 AVC can be remarkably small.

6 REFERENCES

- [1] ITU-T, "Video codec for audiovisual services at p x 64 kbit/s", ITU-T Recommendation H.261, Version 1: November 1990, Version 2: March 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 11172-2 (MPEG-1 Video), March 1993.
- [3] 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), November 1994.
- [4] ITU-T, "Video coding for low bit rate communication", ITU-T Recommendation H.263, Version 1: November 1995, Version 2: January 1998, Version 3: November 2000.
- [5] ISO/IEC JTC 1, "Coding of audio-visual objects - Part 2: Visual", ISO/IEC 14496-2 (MPEG-4 Visual), Version 1: April 1999, Version 2: February 2000, Version 3: May 2004.
- [6] ISO/IEC JTC 1, "Advanced video coding for generic audio-visual services", ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), Version 1: May 2003, Version 2: January 2004, Version 3: September 2004, Version 4: July 2005.
- [7] ISO/IEC JTC 1, "Advanced video coding for generic audio-visual services", ITU-T Recommendation H.264 Amendment 3, ISO/IEC 14496-10/2005: Amd 3 - Scalable extension of H.264 (SVC), July 2007.
- [8] ISO/IEC JTC 1/SC 29/WG 11, "Applications and Requirements for Scalable Video Coding", ISO/IEC JTC 1/SC 29/WG 11, doc. N6052, Brisbane, Australia, 20-24 October 2003.
- [9] ISO/IEC JTC 1/SC 29/WG 11, "Registered responses to the call for proposals on scalable video coding", ISO/IEC JTC 1/SC 29/WG 11, doc. M10569, Munich, Germany, 15-19 March 2004.
- [10] Z. G. Li, S. Rahardja and H. Sun, "Implicit bit allocation for combined coarse granular scalability and spatial scalability" IEEE Transactions on Circuits and Systems for Video Technology, Vol. 16, No. 12, pp.1149-1459, December 2006.
- [11] W. Yao, Z. G. Li and S. Rahardja, "Balanced inter-layer prediction for combined coarse granular scalability and spatial scalability", ISCAS 2008, May 2007.

- [12] H. Schwarz, T. Hinz, D. Marpe and T. Wiegand, "Overview of the scalable H.264/MPEG4-AVC Extension", Proc. of ICIP 2006, Atlanta, USA, 8-11 October 2006.
- [13] M. Wien, H. Schwarz and T. Oelbaum, "Performance Analysis of SVC", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 17, No. 9, pp.1194-1203, September 2007.
- [14] W. Li, "Overview of Fine Granularity Scalability in MPEG-4 Video Standard", IEEE Transactions on Circuits and Systems for Video Technology, Vol.10, No. 5, pp.625-743, August 2000.
- [15] B. G. Haskell, A. Puri and A. N. Netravali, "Digital Video: An Introduction to MPEG-2", New York: Chapman&Hall, 1997.