CODING EFFICIENCY IMPROVEMENT FOR SVC BROADCAST IN THE CONTEXT OF THE EMERGING DVB STANDARDIZATION

Heiko Schwarz [1], *Patrick Ndjiki-Nya* [1], *and Thomas Wiegand* [1,2]

[1]Image Processing Department, Fraunhofer Institute for Telecommunications (HHI), Einsteinufer 37, 10587 Berlin, Germany
[2]Image Communication Group, Technical University of Berlin, Einsteinufer 17, 10587 Berlin, Germany
phone: +49-30-31002-[226|205|617], fax: +49-30-3927200, email: [hschwarz|ndjiki|wiegand]@hhi.fraunhofer.de

ABSTRACT

The Scalable Video Coding (SVC) amendment of H.264/AVC provides benefits for a variety of video applications. One particular interesting application area is the backward-compatible format enhancement in video broadcast. The coding efficiency in broadcast applications is usually limited due to a frequent insertion of random access points. In this paper, we show that the coding efficiency for SVC broadcast can be improved by increasing the interval between enhancement layer random access points. With the introduction of adequate constraints for random access points and a minor adjustment of the decoding process at a channel change, this improvement can be achieved without any impact on the channel change delay. In our experiments, the coding efficiency for spatial scalable coding became virtually identical to that of single-layer coding while providing the same channel change delay characteristics.

1. INTRODUCTION

The SVC amendment [1] of H.264/AVC [2] provides network-friendly scalability allowing partial transmission and decoding of bit streams. Temporal scalability can be efficiently provided using hierarchical prediction structures [3] and did not require any changes to H.264/AVC. Spatial and quality scalability are supported via a layered coding approach. Experimental investigations as the subjective SVC verifications test [4][5] carried out by MPEG showed that the SVC design is capable of providing spatial and quality scalability at the cost of a bit rate increase of about 10% relative to single-layer H.264/AVC coding. The scalability features, the design consistency, the low complexity overhead, and the low coding efficiency degradation in comparison to single-layer H.264/AVC coding make SVC a promising candidate for a variety of video applications.

One particular interesting application area is the backward-compatible format enhancement in video broadcast. Already deployed (single-layer) receivers can still decode the backward-compatible SVC base layer which is delivered in the existing video format, while newly deployed SVC receivers are capable of decoding the new enhanced video format. Targeting this application area, the Digital Video Broadcasting (DVB) consortium has recently extended its specification TS 101 154 [6] to support SVC and is currently working on a corresponding extension of the specification TS 102 005 [7]. The new version of TS 101 154 and the draft for TS 102 005 envisage the support of SVC for transmission systems based on the MPEG-2 Transport Stream and the Internet Protocol. Depending on the application capabilities the Scalable Baseline or Scalable High profile is supported. The Scalable High profile is supported up to Level 4.2, which enables the introduction of the improved 1080p50/60 HDTV format while retaining full compatibility with existing 720p50/60 or 1080i25/30 HDTV receivers. For mobile DVB-H applications, a VGA service could be introduced in addition to today's QVGA services.

In broadcast video services, so-called Random Access Points (RAPs), i.e., pictures at which a decoder can start decoding the bit stream, must be provided in regular intervals in order to ensure an acceptable behaviour at channel changes. Since the picture of a RAP must be usually intra-coded without prediction from previously transmitted pictures, this frequent provision of RAPs usually limits the coding efficiency. In this paper we will show that an increase of the interval between enhancement layer RAPs can improve the coding efficiency for SVC broadcast without any impact on the channel change delay. The proposed approach has been included in TS 101 154 as an informative Annex.

An overview of the general concept is given in the next section. Sec. 3 introduces constraints for RAPs in SVC broadcast bit streams that allow starting the decoding of an SVC bit stream at any base or enhancement layer RAP with only a minor adjustment of the decoding process, which is described in Sec. 4. Experimental results demonstrating the achievable coding efficiency gains are presented in Sec. 5.

2. OVERVIEW

In single-layer H.264/AVC coding, random access is always enabled at IDR (Instantaneous Decoding Refresh) pictures, which basically reset the decoder status. However, non-IDR intra pictures with additional constraints are usually sufficient and the frequent insertion of IDR pictures decreases the coding efficiency for high-delay prediction structures. This issue, which is similar to the well-known open GOP / closed GOP problem for MPEG-2 Video [8], is briefly discussed on the basis of Figure 1. When inserting IDR pictures as RAPs as shown in Figure 1(a), the sequence of pictures consisting of the IDR RAP (black box) and all following pictures in decoding order (shaded boxes) can be decoded independently of all pictures that precede the RAP in decoding order (white boxes). However, the relative number of bi-directionally predicted pictures is reduced and the coding efficiency is degraded. Additionally, temporal blocking effects may be visible due to the break in the prediction chain. By using non-IDR intra pictures as RAPs as illustrated in Figure 1(b), this coding efficiency loss is avoided. With the constraint that all pictures that succeed the RAP in output order must not reference any picture that precedes the RAP in output order for inter prediction, the RAP (black box) and all pictures that succeed the RAP in output order (shaded boxes) can be decoded independently of all pictures that precede the RAP in decoding order (white boxes). When accessing the bit stream at such a RAP, the pictures that succeed the RAP in decoding but precede it in output order (grey boxes) cannot be decoded and are ignored.

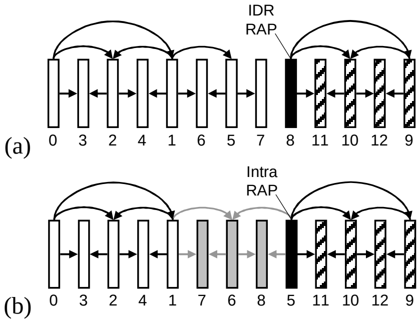

Figure 1 – Random access points in coding structures with hierarchical B pictures: (a) using IDR pictures, (b) using non-IDR intra pictures. The pictures are depicted in output order; the numbers below the pictures indicate the decoding order.

H.264/AVC provides the recovery point SEI message for signalling RAPs, where it is also possible to specify that the recovery point (i.e., the picture for which the picture itself and all following pictures in output order are correct or approximately correct) is different from the RAP. Mechanisms for RAP signalization are also supported in transport protocols such as MPEG-2 Systems [9]. In the following description, we will concentrate on the case that RAPs are intra pictures and represent recovery points, since this is often required in broadcast applications (cp. TS 101 154).

The common approach for providing RAPs in SVC broadcast bit streams is to regularly insert pictures for which all layer pictures are intra coded (the enhancement layer pictures can still use inter-layer intra prediction) and represent recovery points as described above. This ensures that the channel change delay characteristic is basically the same for all layers. However, the concept of generalized IDR pictures in SVC [10] can be extended to RAPs. With the SVC design it is possible to insert RAPs for different layers at different time instances. The decoding of a particular layer can only be started at a RAP for this layer. For a picture that represents a RAP for a particular layer, the corresponding layer picture must be intra-coded, but all other layer pictures can be inter-coded using previous pictures as references, which increases the coding efficiency for these layers.

Since all layers in an SVC bit stream represent the same content and the decoding of this content can be basically started at any base or enhancement layer RAP, the intervals between RAPs for the enhancement layer could be increased relative to the intervals between base layer RAPs. This usually improves the enhancement layer coding efficiency, since more enhancement layer pictures can be inter-coded, but it does not necessarily increase the channel change delay for enhancement layer decoders. The decoding can either be normally started at an enhancement layer RAP or it can be started at a base layer RAP. In the latter case, the decoder would start the decoding of the base layer and switch to the enhancement layer at the next enhancement layer RAP. In order to ensure a seamless video playback, the decoded base layer pictures are upsampled to the enhancement layer format. The only drawback of this approach is that the user may see a quality change in the displayed video at the point of layer switching. But when the interval between enhancement layer RAPs is kept reasonably small, this quality change is usually not disturbing, since it appears in the time period after the channel change in which the users eye adjusts to the new video content. In Sec. 4 we will highlight a possibility to further reduce the visibility of such quality changes.

3. SVC RANDOM ACCESS POINTS

When all RAPs provided in an SVC bit stream represent IDR pictures, the potentially required switching from base to enhancement layer decoding after random access is straightforward; the IDR picture generalization in SVC [10] was particularly designed for such a layer switching. For non-IDR RAPs, the switching from a lower to a higher layer can be more complicated due to interdependencies between base and enhancement layer coding. This can be illustrated using the simple example in Figure 1(b) and assuming a two-layer SVC bit stream with the same coding structure in base and enhancement layer. When we switch from base to enhancement layer coding at the depicted RAP (black box), we cannot decode the enhancement layer representations of the pictures that follow the enhancement layer RAP in decoding order but precede it in output order (grey boxes) due to missing references for inter prediction. In order to enable seamless video playback, we have to decode the base layer representations for these pictures. But these base layer representations depend on the base layer representation of the enhancement layer RAP, for which we would have to decode both base and enhancement layer. For more complicated prediction structures, it may be even necessary to decode and store base and enhancement layer representations for more than one picture, which would significantly increase the decoder complexity and the memory requirement.

In order to enable a seamless switching between base and enhancement layer with a negligible increase in decoder complexity, we introduce the following constraints:

1. When a picture represents a RAP for a particular layer, the layer picture for this layer must be intra-coded.
2. Enhancement layers of pictures that follow a RAP in output order must not reference any picture that precedes the RAP in output order through inter prediction.
3. Pictures that succeed a RAP in output order shall not be transmitted before the RAP or any picture that precedes the RAP in output order.
4. When a picture represents a RAP for a particular layer, it must also represent a RAP for all lower layers.
5. Each RAP must have temporal_id equal to 0.

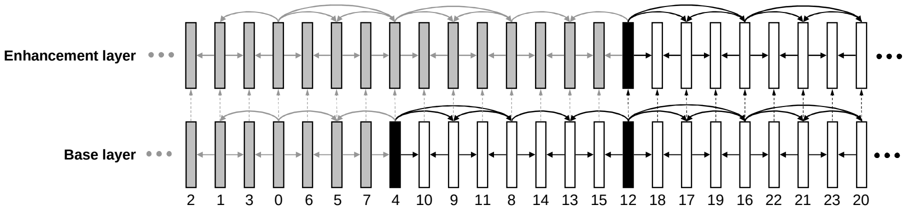

Figure 2 – Example for the random access with switching from base to enhancement layer in a coding structure with hierarchical B pictures. The pictures are depicted in output order; the numbers below the pictures indicate the decoding order.

On the one hand, these constraints allow starting the decoding of an SVC bit stream at an arbitrary RAP, including the switching from a lower to a higher layer, with a negligible decoder complexity overhead. And on the other hand, they still allow employing high efficient prediction structures such as hierarchical B pictures [3], which have been proven to be the most efficient known coding structure for high-delay applications. In comparison to the usual RAP constraints for broadcast (cp. Sec. 2) we have tightened condition 2, which now completely forbids inter prediction across RAPs in order to ensure that for each picture, except the RAP at which the layer switching is done, only a single representation has to be decoded and stored. Condition 3 ensures that the decoder can switch between base and enhancement layer decoding at a single well-defined point, which is important because a multiple switching would require the operation of two decoded picture buffers. Condition 4 targets a further restriction of the decoder complexity overhead. Since the intra macroblocks in the lower layer have to be decoded anyway for inter-layer prediction of the enhancement layer, the overhead for decoding the lower layer representation of the RAP is negligible. Finally, condition 5 was introduced for ensuring that the same RAPs are available for all temporal sub-streams that are present in a scalable bit stream.

4. DECODING PROCESS AT RANDOM ACCESS

In the following, we describe the modification of the decoding process at random access. For keeping the description simple, we restrict it to spatial scalable coding with a base and a single enhancement layer. A more general specification can be found in Annex G of TS 101 154 [6].

After a channel change, an SVC receiver starts the decoding of the SVC bit stream at the first RAP it receives. If this RAP is an enhancement layer RAP, the decoding starts as specified in the standard [2] and no further adjustments are required. If however, the first received RAP represents only a base layer RAP, the start of the decoding process should proceed as specified in the following ordered steps:

1. Decode all base layer representations of the RAP and all following pictures that precede the next enhancement layer RAP in decoding order.
2. If the enhancement layer RAP represents an IDR picture for the enhancement layer, the enhancement layer representation is decoded and inserted in the decoded picture buffer (DPB) as specified in the standard.
3. If the enhancement layer RAP does not represent an IDR picture for the enhancement layer, the following ordered steps are performed:
   a. For the enhancement layer RAP, decode both the base and enhancement layer representation. The base layer representation is normally stored in the DPB, while the enhancement layer representation is stored in a temporary buffer outside of the DPB.
   b. Decode the base layer representations for all pictures that follow the enhancement layer RAP in decoding order, but precede it in output order.
   c. Mark all base layer representations as "unused for reference" and insert the temporary stored enhancement layer representation in the DPB.
4. Continue the enhancement layer decoding as specified in the standard for all pictures that follow the enhancement layer RAP in both decoding and output order.

In addition, for all pictures for which only the base layer representation is decoded, this representation is upsampled to the enhancement layer video format before displaying it in order to ensure a seamless video playback.

The process of accessing an SVC bit stream at a base layer RAP is further explained based on the example in Figure 2, which shows a two-layer SVC bit stream using hierarchical B pictures with 3 dyadic temporal levels. The RAPs, which are all non-IDR intra pictures, are represented by black boxes and non-decoded representations are represented by grey boxes. The first picture that is received after a channel change is picture 0. This picture and the following pictures 1 to 3 don't represent RAPs and are completely ignored. The decoding process starts with picture 4, which represents a base layer RAP. The following pictures 5 to 7 are ignored, since they precede the RAP in output order. The base layer decoding is normally continued with pictures 8 to 11 until the first enhancement layer RAP in picture 12 is received. For this picture both the base and enhancement layer representation is decoded. The base layer representation is normally inserted in the DPB, while the enhancement layer representation is stored in a temporary buffer. For the following pictures 13 to 15 that precede the enhancement layer RAP in output order, the normal base layer decoding process is continued. Then, the decoding process switches from base to enhancement layer decoding. All base layer representations in the DPB are marked as "unused for reference" and the temporary stored enhancement layer representation for picture 12 is inserted in the DPB. Now, the normal enhancement layer decoding process as specified in the standard is continued starting with picture 16. For the pictures 4, 8 to 12, and 15 to 17, the decoded base layer representations are upsampled to the enhancement layer video format before they are displayed.

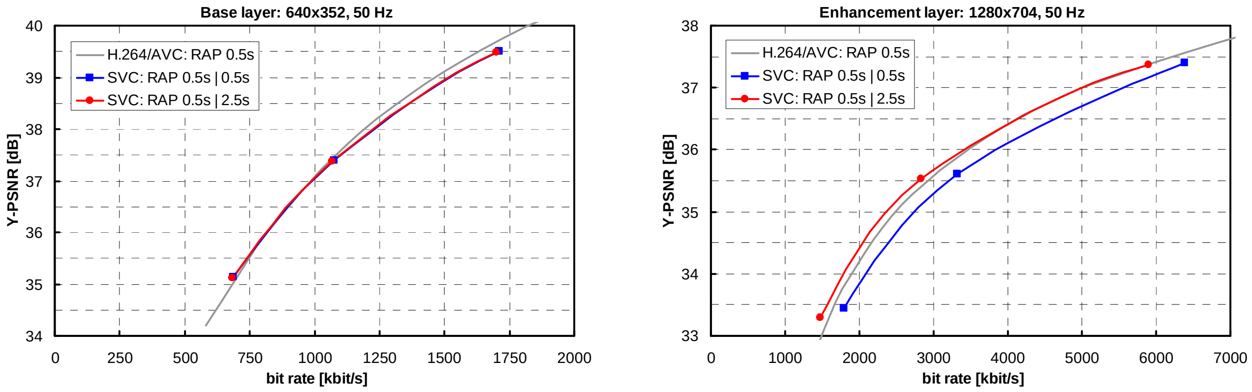

Figure 3 – Simulation results comparing the coding efficiency of the different configurations for the City test sequence.

It should be noted again that the additional decoding of the base layer representation for the switching point (picture 12 in the example) does not increase the decoder complexity, since the base layer representation is completely intra-coded (cp. condition 4 in Sec. 3) and the reconstructed intra macroblocks are required anyway for inter-layer prediction of the enhancement layer pictures. The only minor drawback is that we need an additional frame store for temporarily storing the enhancement layer reconstruction of the switching point.

The switch between base and enhancement layer quality always occurs at a single point in output order (the first received enhancement layer RAP). When keeping the interval between enhancement layer RAPs reasonable small, this quality change is usually not disturbing, since it occurs in the time period after channel switching in which the users eye is still adjusting to the new video content.

The visibility of the quality change can further be reduced by applying a time-varying low-pass filter (before display) to the first pictures for which the enhancement layer representation is displayed. The cut-off frequency of the low-pass filter is continuously increased in output order. For the example in Figure 2, this would mean that we low-pass filter the reconstructed picture 12 with a cut-off frequency that is selected according to the ratio between the base and enhancement layer picture sizes. For the following pictures in output order (i.e., pictures 18, 17, 19, 16, etc.), the cut-off frequency is continuously increased until the enhancement layer pictures are displayed without the low-pass filtering.

5. EXPERIMENTAL RESULTS

For demonstrating the effectiveness of proposed approach of increasing the interval between enhancement layer RAPs, we compared it to the common method of providing base and enhancement layer access point in the same regular intervals as well as to single-layer H.264/AVC coding. For the SVC configurations we used the Scalable High profile with one enhancement layer. The single-layer references were coded conforming to the High profile. We used the same coding structure with hierarchical B pictures and 4 dyadic temporal levels (GOP size of 8) for all configurations. All encoding runs including the single-layer runs were performed using the same software (based on the JSVM reference software, version 8.5) and a similar degree of encoder optimizations. For the SVC bit streams we additionally employed the multi-layer encoding concept presented in [11] in order to distribute the coding efficiency losses relative to single-layer coding between base and enhancement layer. The quantization parameters (QP) were set constant for each encoder run and no rate control algorithm was employed. In order to obtain a reasonable bit rate distribution between base and enhancement layer, the enhancement layer QP was set equal to $\mathrm{QP}_\mathrm{B}+4$ with QP_B representing the base layer QP. We run simulations for 6 sequences with a resolution of 1280x704 luma samples and a frame rate of 50 Hz (similar to the 720p50 format). The base layer was encoded with a resolution of 640x352 luma samples and the same frame rate.

Only the first picture of each sequence was coded as IDR picture; all other RAPs were encoded as non-IDR intra pictures. In the single-layer configuration, RAPs were inserted every 24 pictures (i.e., about every half second). For SVC, two different RAP configurations were used. In the first configuration, which represents the common method for providing random access points, RAPs for base and enhancement layer were inserted every 24 pictures as in the single-layer configuration. In the second configuration, which represents the proposed approach, base layer RAPs were again inserted every 24 pictures, but enhancement layer RAPs were inserted only every 120 pictures (i.e., about every 2.5 seconds). It should be noted that all bit streams are associated with the same channel change delay characteristics assuming that SVC receivers use a decoding process similar to the one described in Sec. 4.

As an example, the obtained rate-distortion curves for the City sequence are shown in Figure 3. It can be seen that the base layer coding efficiency for both SVC configurations is virtually the same and very close to that of single-layer coding. For the enhancement layer, the coding efficiency of SVC could be significantly increased by enlarging the interval between enhancement layer RAPs and it actually became even slightly better than that of single-layer H.264/AVC coding while providing the same channel change delay.

A similar behaviour was observed for all tested sequences. The modification of the enhancement layer RAP interval doesn't have any noticeable impact on the base layer coding efficiency (a very small influence is a consequence of the multi-layer encoder control [11]); the average rate overhead for the base layer in comparison to single-layer coding was about 1 to 5 %. The average results for the enhancement layer coding are summarized in Table 1. The proposed increase of the enhancement layer RAP interval results in a bit rate saving of about 10 % for the enhancement layer resolution relative to the common approach of providing RAPs in SVC broadcast bit streams. This rate saving lead to an overall coding efficiency for SVC broadcast bit streams that is virtually identical to that of single-layer H.264/AVC coding with the same channel change delay.

Table 1 – Summary of experimental results.

<table>
<tr><th rowspan="2">test sequence</th><th colspan="2">bit rate overhead of SVC relative to single-layer coding</th><th rowspan="2">bit rate saving with increased RAP interval</th></tr>
<tr><th>RAP 0.5s | 0.5s</th><th>RAP 0.5s | 2.5s</th></tr>
<tr><td>Aloha Wave</td><td>9.6 %</td><td>0.6 %</td><td>−9.0 %</td></tr>
<tr><td>Big Ships</td><td>6.9 %</td><td>−0.6 %</td><td>−7.5 %</td></tr>
<tr><td>City</td><td>8.7 %</td><td>−3.5 %</td><td>−12.2 %</td></tr>
<tr><td>Dancer</td><td>9.8 %</td><td>7.6 %</td><td>−2.2 %</td></tr>
<tr><td>Old Town Pan</td><td>15.7 %</td><td>−7.8 %</td><td>−23.4 %</td></tr>
<tr><td>Sailormen</td><td>6.2 %</td><td>1.8 %</td><td>−4.5 %</td></tr>
<tr><td>average</td><td>9.5 %</td><td>−0.3 %</td><td>−9.8 %</td></tr>
</table>

In order to test the effect of applying a time-varying low-pass filter after a switch to enhancement layer coding, we simulated a low-pass filter with adjustable cut-off frequency by the following approach. A decoded enhancement layer picture is downsampled (using the non-normative JSVM downsampling filter) to an intermediate resolution of AxB luma samples and this version is then upsampled to the original enhancement layer resolution (using the SVC upsampling filter). The strength of the resulting low-pass filter is determined by the intermediate resolution. The time-varying low-pass filter was simulated by linearly increasing the intermediate resolution from the base to the enhancement layer resolution over a certain period of time after a switch from base to enhancement layer coding. Informal tests showed that the quality change at a switching point becomes nearly invisible when the transition period is about one second or larger. Figure 4 illustrates the effect of the low-pass filtering on the PSNR of the displayed pictures after a channel change. The transition interval was set equal to 1 second.

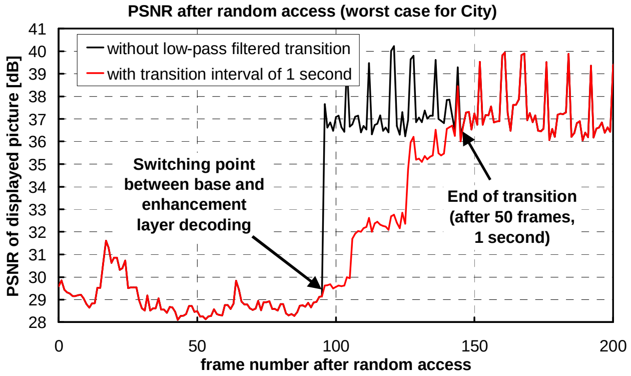

Figure 4 – Example for illustrating the impact of the low-pass filtering on the PSNR of the displayed pictures after random access with a switching from base to enhancement layer decoding.

6. CONCLUSION

It was shown that the coding efficiency of SVC broadcast bit streams can be significantly improved by increasing the interval between enhancement layer RAPs. With the introduction of suitable constraints for RAPs in SVC broadcast bit stream, this coding efficiency improvement doesn't have any impact on the channel change characteristics and does require only a minor adjustment of the decoding process at random access. The decoding complexity is basically not increased, but one additional frame store is needed. Although a quality change might occur in the displayed video just after channel change, we found that this quality change is generally not disturbing when the interval between enhancement layer RAPs is kept reasonably small and we highlighted an approach for further reducing this effect. Our experimental results indicate that two-layer spatial scalability in broadcast applications could be provided at virtually the same bit rate that would be required for single-layer coding of the enhancement layer resolution without any compromise in reconstruction quality or channel change delay.

REFERENCES

[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. CSVT*, vol. 17, no. 9, pp. 1103-1120, Sep. 2007.

[2] ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), "Advanced video coding for generic audiovisual services," version 11, 2009.

[3] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MCTF," *Proc. ICME'06*, Toronto, Canada, pp. 1929-1932, July 2006.

[4] ISO/IEC JTC 1/SC 29/WG 11, "SVC Verification Test Report," Doc. N9577, Antalya, Turkey, Jan. 2007.

[5] T. Oelbaum, H. Schwarz, M. Wien, and T. Wiegand, "Subjective performance evaluation of the SVC extension of H.264/AVC," *Proc. ICIP'08*, San Diego, CA, USA, pp. 2772-2775, Oct. 2008.

[6] ETSI TS 101 154, "Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcasting Applications based on the MPEG-2 Transport Stream", ver. 1.9.1, 2009.

[7] ETSI TS 102 005, "Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in DVB services delivered directly over IP protocols", ver. 1.3.1, 2007.

[8] ITU-T Rec. H.262 and ISO/IEC 13818-2 (MPEG-2 Video), "Generic coding of moving pictures and associated audio information: Video," 1994.

[9] ITU-T Rec. H.222.0 and ISO/IEC 13818-1 (MPEG-2 Systems), "Generic Coding of moving pictures and associated audio information: Systems," 2007.

[10] Q. Shen, H. Li, Y.-K. Wang, and M.M. Hannuksela, "Enhancement layer IDR (EIDR) picture," *Joint Video Team*, Doc. JVT-R053, Bangkok, Thailand, Jan. 2006.

[11] H. Schwarz and T. Wiegand, "R-d optimized multi-layer encoder control for SVC," *Proc. ICIP'07*, San Antonio, TX, USA, pp. 281-284, Sep. 2007.