Scalable Video Coding Extension of H.264/AVC

Anamaria Bjelopera [1], Sonja Grgić [2]
[1] University of Dubrovnik, Department of Electrical Engineering and Computing, Ćira Carića 4, 20000 Dubrovnik, Croatia
[2] University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
anamaria.bjelopera@unidu.hr

Abstract - SVC (scalable video coding) was introduced as an amendment of H.264/AVC by Moving Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG) in 2007. SVC can provide different types of scalability and retain reconstruction quality characteristics similar to single layer H.264/AVC but in trade of higher complexity of the decoder compared to H.264/AVC. This paper presents an overview of improved H.264/AVC standard and describes tools for supporting temporal scalability, spatial and SNR (signal-to-noise ratio) scalability. Temporal scalability is realized by the hierarchical B-frame (bi-predictive picture) structure which shows better results than structures used in prior video coding standards. Spatial scalability uses inter-layer prediction from any other layer and residual prediction which shows better results in coding efficiency. For quality scalability different methods are analyzed and the results are compared. The paper illustrates the results for coding efficiency with concept of using combined types of scalability.

Keywords **-** H.264/AVC, SVC, scalability, enhancement layer

I. INTRODUCTION

H.264/AVC [1-2] is one of the most important video coding standards which is used in many different applications such as multimedia applications, mobile TV, Internet video streaming and HDTV broadcasting so the video signal could be displayed on different equipment with range of sizes. This standard enables cooperation of different products with different services of applications. The advantage of H.264/AVC is reduction of bit rate and improvement of coding efficiency [3]. Apart from having good characteristics, there was the need for SVC [3-4] because of the technology of receiving equipment that was changing every day and different transmission conditions that were varying in connection quality.

Scalability is referred to removing parts of the video bit stream that would be suitable for different user and network capabilities. Video bit stream is scalable when parts of the stream can be removed [3]. The rest of the stream also builds a valid substream for target coder which has lower quality than the original stream. The minimum bit stream that can be decoded is called base layer and the rest of the bits are called enhancement layers which give more details in reconstruction of video. If bit streams do not allow removing parts of bit stream, they are called single-layer bit streams.

SVC gives a number of possibilities in different applications. For instance, there are heterogeneous devices and environment which are receiving same video content [5]. They need adaptation of once-encoded content which is suitable for their capabilities (different bit rate, frame rate, picture size). Source content is once-coded for highest resolution and bit rate, so that devices with restricted needs can decode only parts of bit stream which they are going to use. Some of applications need only parts of bit stream so terminals do not need to decode the whole bit stream. It also contains parts of different decoded video quality [3]. Stronger protection can be used on video content with higher importance and therefore resilience on different errors generated in video transmission conditions can be improved. Using SVC in applications where video content has to be stored, the parts of stream with high quality can be deleted after some time, and those with lower quality can be stored for long time.

Scalability has already been provided with usage of scalable profiles but not in terms of picture size and quality. These types of scalability come in addition of higher decoder complexity and reduced coding efficiency. Used types of scalability are spatial, quality and temporal scalability shown in Fig. 1 [6]. Temporal scalability reduces frame rate of the original bit stream (temporal resolution) and spatial scalability reduces the size of picture (spatial resolution). Quality scalability provides substreams with same temporal and quality resolution as the original bit stream but the SNR is lower. This type of scalability is also called fidelity or SNR scalability. SVC bit stream can be used with different combinations of temporal, spatial and SNR scalability. SVC supports different resolution ratios that can be changed at any point in time meaning the size of the original bit stream and substreams are not restricted.

Each SVC stream contains a substream which is compatible with non-scalable profile of H.264/AVC. Also, every type of application has their own demands on which is based the trade between the scalability and coding efficiency.

This paper shows overview of SVC extension of H.264/AVC. Section II describes basics of architecture of H.264/AVC. Section III presents different types of scalability including temporal, spatial and SNR scalability and their results regarding coding efficiency. Section IV analyses results of combined types of scalability. Section V concludes the paper.

II. BASIC ARCHITECTURE OF H.264/AVC

NAL (network abstraction layer) and VCL (video coding layer) [3, 7-8] are basic components of designing SVC. VCL implicates coded source data and NAL gives them header information adequate for different type of systems and

conditions of transport. NAL units are contained of coded video data and each of them presents a packet of integer number of bytes. The first byte of NAL unit presents the header of each unit which indicates the type of data in that unit and the rest of bytes contain payload data of the same type. The encoder generates a stream called a NAL unit stream.

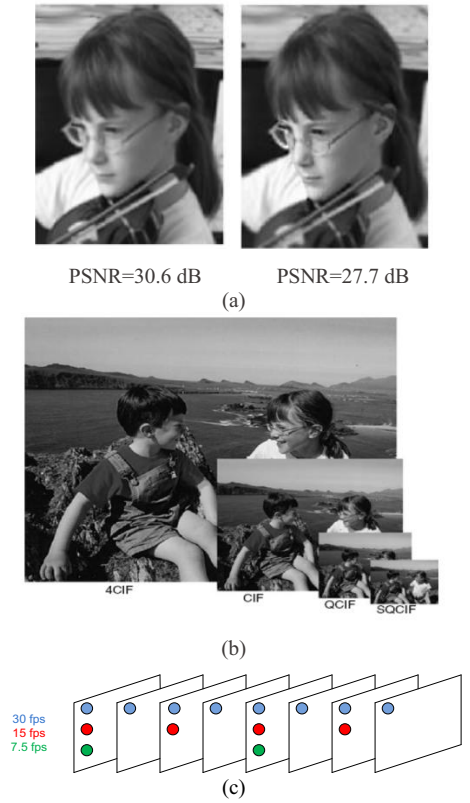


Fig. 1 Different types of scalability: (a) quality scalability, (b) spatial scalability, (c) temporal scalability [6]

NAL units are sorted in VCL NAL units which contain coded video data of picture samples and non-VCL NAL units which ensure associated additional information like parameter sets that are changeable for each video sequence and data that are not necessary for decoding but can facilitate it. A specific set of NAL units forms an access unit. The decoding process of access unit results in one decoded picture. A set of sequential access units presents coded video sequence. First access unit in every video sequence is called IRD (instantaneous decoding refresh) and it is always coded as I-picture [3]. It is used for current refreshing of decoding process. A NAL unit stream contains one or more coded video sequences. Each decoding process of coded video sequence is independent of decoding of any other coded video sequence. A scalable bit stream can be reduced in terms of spatio-temporal resolution or SNR scalability by discarding NAL units.

VCL has block-based hybrid video coding which enables better compression efficiency and also better flexibility than previous standards [3]. Each part of bit stream presenting video source with different spatial resolution and fidelity is called layer. Each layer has its own layer identification. The layer whose data are used to predict data in other layers is called reference layer. The layer which has layer identifier equal to zero is referred to as base layer and it is not present in every access unit. Layers which use data of other layers are called enhancement layers. There are two types of enhancement layers. The one that changes resolution in relation with reference layer is called spatial enhancement layer and the one that has the same spatial resolution as reference layer is called fidelity enhancement layer.

There can be 128 layers in a bit stream [7]. The number of layers varies based on different application requirements. Encoding process of each layer occurs separately for each one of them. Prediction of layers is achieved by spatial up-sampling of lower layer pictures or pictures of the same layer that are adjacent in time. Each layer input picture is divided into smaller units such as macroblocks and slices. Each macroblock is a quadratic area of 16x16 samples of luma component and 8x8 samples of two chroma components with the use of 4:2:0 sampling format. The macroblocks are organised into slices. Each macroblock can be temporally or spatially predicted.

A. Profiles and levels

Profiles define groups of applications with similar requirements [3]. It defines a set of coding tools used for generating a bit stream. There are three profiles: the Scalable Baseline, the Scalable High and the Scalable High Intra Profile. The Scalable Baseline profile is used for mobile broadcasting and conversational applications with lower complexity of coding. The Scalable High is created for broadcasting, streaming and storage applications. The Scalable High Intra profile is used for professional applications.

III. TYPES OF SCALABILITY

The most important requirement for scalable video coding standard to become successful is coding efficiency and complexity in terms that new tools should be added only if they are needed for this kind of requirements.

A. Temporal Scalability

If the set of access units of a bit stream is divided into set of temporally based layers it is called temporal scalability. The layers are consisted of one base layer with several temporal enhancement layers [3]. The counting of layers is identified with temporal layer identifier T which starts with 0 referred to as base layer and increasing by one for the next enhancement layers. Hybrid video coding can be enabled by using motion-compensated prediction with reference pictures that have temporal layer identifier less or equal to that of picture to be predicted. Also, every picture can be used as reference picture and used for prediction. Generally, temporal scalability enables one bit stream to uphold multiple frame rates. These multiple frame rates are enabled by using different prediction structures.

Hierarchical prediction structures are enabling better coding standards like MPEG-2/4. Fig. 2 illustrates coding

structure for temporal scalability called hierarchical B structure. GOP (group of pictures) is consisted of one key picture and multi-level B-picture decomposition between two key pictures [10]. Key pictures are part of the lowest temporal layer and they use previous key pictures as reference pictures. Key frames are usually P (predicted picture) or I (intra-coded picture) frames. Enhancement layer pictures are coded as B-pictures. Lower layer pictures are coded before those of higher layers. The structure of prediction is usually dyadic. This B structure can also be realized using P slices rather than B-slices.

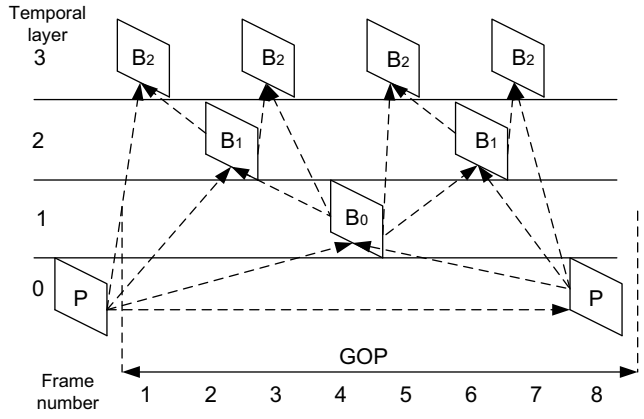


Fig. 2 Hierarchical B frame structure with GOP of 8 pictures [9]

Coding efficiency can be improved by controlling quantization parameter (QP) [9]. Each layer has different QP. The lowest QP-s are used for coding key pictures and therefore these pictures have the best quality simply as they are used as prediction pictures for other temporal layer pictures. Higher temporal layers have increased QP values and

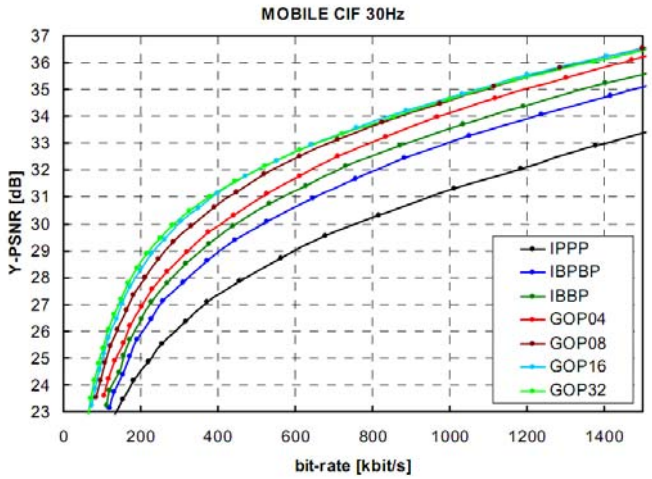


Fig. 3 Coding efficiency comparison of different of hierarchical B-pictures and classical coding structures for the sequence "Foreman" [10]

pictures of these layers have lower quality and smaller quantity of bits. The reason for increasing QP on higher levels is that quantization noise which is lower in lower temporal layers should not be coded again in higher temporal layers. Motion-compensation prediction also results in pictures that have better quality because of using key pictures with higher quality.

Fig. 3 depicts results in PSNR measurements for different prediction structures of "Mobile" sequence in CIF resolution and a frame rate of 30 Hz which is presented with high spatial detail. It compares dyadic hierarchical prediction structures with the most common used prediction structures as IPPP, IBBP. PSNR gains are around 1 dB compared to IBBP coding structure. The coding efficiency can be improved by enlarging the GOP size and the best is realized for GOP between 8 and 32 pictures.

B. Spatial Scalability

SVC applies spatial scalability by decomposing the original video into number of layers as in MPEG-4 video coding standard [3, 7-8, 11]. Temporal and spatial predictions of each spatial layer are applied on reference pictures at that particular spatial layer.

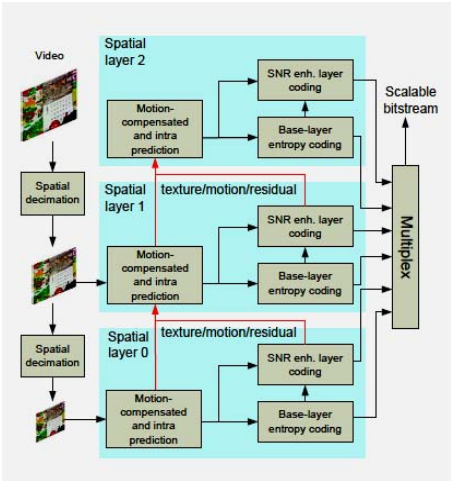


Fig. 4 SVC encoder structure with three spatial layers [11]

Fig. 4 shows the encoder structure of SVC encoder which operates with three spatial layers. It has pyramid coding technique which divides the original video into a number of pyramid videos. The frames are predicted as in H.264/AVC to generate the base layer and redundancy between layers is applied by inter-layer prediction. Aspects of this prediction imply the use of intra, motion and residual data. Information of base layer is than coded with SNR enhancement layer coding block to create quality scalable stream. Multiplex block is going to create a single scalable bit stream consisted of all the layers multiplexed together.

Unlike the MPEG-4/2 which uses inter-layer prediction of closely situated spatial layers, SVC applies inter-layer prediction from any other layer. To improve coding efficiency SVC allows using statistical similarities between information of lower layers [8].

Fig. 5 shows spatial scalability with three layers: base layer $\mathbf{S_0}$ and two enhancement layers $\mathbf{S_1}$ and $\mathbf{S_2}$. There is a different number of pictures in this figure and there is a combination of temporal and spatial scalability. The vertical

arrows show the inter-layer prediction where particular layers use information of other layers. The redundancy is reduced and the coding efficiency is better [8].

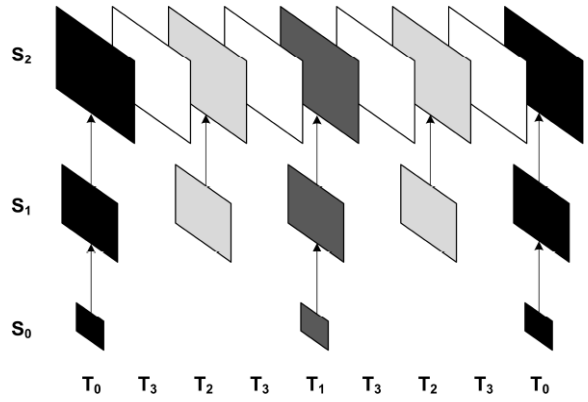


Fig. 5 Spatial scalability with three layers [8]

In MPEG-4 the data of reconstructed lower layers are upsampled and motion predicted but this kind of information is not the most appropriate for inter-layer prediction. There are two kind of prediction which show better results in coding efficiency: residual prediction and prediction on a macroblock or submacroblock level. SVC can also use inter-layer or intra-layer prediction [3].

Intra-layer prediction can predict samples out of spatially predicted neighboring macroblocks of same picture of the same layer or spatially displaced macroblocks between pictures of the same layer. The difference between the original and prediction signal is transformed into frequency domain and each value of transform coefficients is reduced and then quantized. At the end these coefficients are entropy coded. Motion vectors are predicted based on motion vectors of adjacent previously coded blocks. The difference between predicted motion vector and current motion vector is coded and then transmitted. The residual signal is decoded by inverse scaling and inverse transformation of each transform coefficient and added to prediction signal.

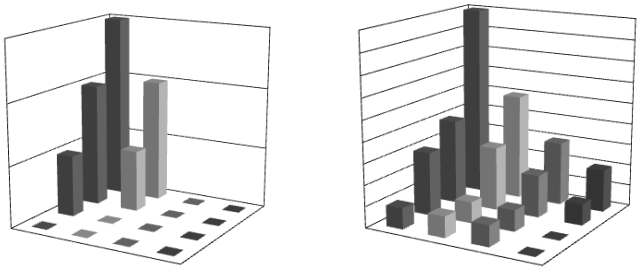
SVC coder can choose the type of prediction whether it would be inter-layer or intra-layer prediction. The layer used for inter-layer coding is called reference layer. There is a coding mode where macroblocks of signal are predicted by using collocated blocks in reference layer without any other supplemental information such as motion parameters. The collocated blocks of reference layer are intracoded and the reconstructed signal is up-sampled. This is called inter-layer intra prediction.

Inter-layer motion prediction reduces redundant data (motion vectors) from motion information and it is called base-layer mode. There should be at least one collocated block that is not intracoded. Macroblocks are predicted out of collocated blocks of reference layer. Also, for other motion parameters (motion vectors) this mode uses information of reference layers. There is a motion prediction flag which indicates usage of inter-layer motion vector predictor. There are two values of motion prediction [3]. Flag 1 indicates the usage of motion vectors predictors derived out of collocated macroblocks in the reference layer. Flag 0 specifies using proper motion vectors of corresponding blocks from enhancement layer and the prediction of motion vectors is specified as in H.264/AVC standard.

Inter-layer residual prediction is reducing the bit rate of sent information. The information are regarding to residual signal. The residual, meaning the difference between collocated blocks of reference layer, is upsampled and used for prediction of residual signal of enhancement layer. The difference signal has less energy than the original signal and is coded with transform coding.

C. Quality Scalability

Quality scalability is similar to spatial scalability [3, 8, 12]. It uses same sizes of pictures for both base and enhancement layers. This method is called coarse-grain scalability (CGS) and shown in Fig. 7 (a). It divides video into a few quality layers and uses spatial scalability methods of coding but it does not use upsampling. The prediction of interlayer intra signals and residual signals is achieved in domain of transforming coefficients so that the decoding on the decoder side can be simplified. Transform coefficients are quantized and then encoded. Quality scalability makes better image quality with reducing the quantization steps which is illustrated in Fig. 6. There are more quantization levels and there is more refined transition between different transform coefficients. Hence, there are more details in pictures so the quality of pictures is also better.

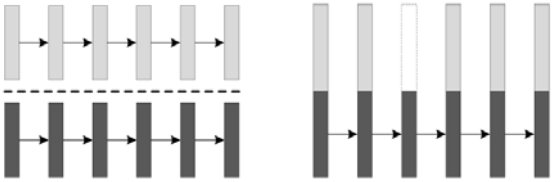


(a) Four Quantization Levels (b) Ten Quantization Levels

Fig. 6 Quantization of Transform Coefficients [8]

There are only a few bit rates that can be used in this type of scalability. The disadvantage of this approach is existence of as many rate points as there are number of layers. The difference between bit rates of consecutive layers is ought to be bigger for better coding efficiency. Motion prediction is performed in each spatial layer separately.

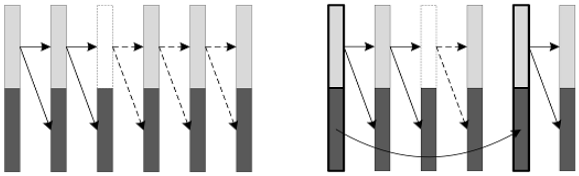
There is another possibility of quality scalability that is better than CGS. It is called FGS (fine grain scalability) and illustrated in Fig. 7 (b). Motion compensation of FGS is realized on the lowest quality level. It uses always available base layer for reconstruction. Motion compensation loop is not affected by the lost of packets with better quality. The bit rate of video stream can be reduced by omitting some packets of enhancement layers. The coding efficiency is reduced because the information of quality enhancement layers is not used for motion prediction.



(a) CGS (b) FGS

Fig. 7 Coarse grain scalability and fine grain scalability [8]

Method that reuses more information from reference pictures by implementing motion prediction in higher quality layers than those used in e.g. FGS is shown in Fig. 8 (a).



(a) drift (b) medium grain

Fig. 8 Different types of quality scalability [8]

The video stream is susceptible to drift. Drift is a case of decoder and encoder working without synchronization and using different reference images. The decoder can not detect losing some packets in transmission of enhancement layer packets and uses lower quality reference picture for motion compensation as the encoder. Fig. 8 (a) performs motion estimation in the enhancement layer at the encoder. At the third frame only the base layer reaches the decoder. Fourth frame uses incomplete reference picture, making decoded video damaged. The drift effects accumulate over time which results in quality loss.

MGS (medium grain scalability) is another method of quality scalability which uses the concept of key pictures shown in Fig. 8 (b) that use only base layer for prediction. Every picture has a flag which shows whether motion prediction exploits the base layer reconstruction or enhancement layer reconstruction of reference picture. Hence, it can choose which quality layer is used for prediction so that prediction can be derived in enhancement layer with updates in the base layer at key pictures. Fig. 8 (b) shows that every fourth picture is used as key picture and motion prediction is updated in the base layer. There is a better control of drift effects.

Fig. 9 compares coding efficiency of CGS and FGS strategies to single-layer coding for the hierarchical-B structure with a GOP size of 16 pictures. The QP difference of the lowest and the highest SNR layer is 12. CGS is used with different delta QP (DQP=6 and DQP=2) which represents the difference of QP of two sequential layers. The diagram shows a black curve which represents single-layer coding. The blue and the green curve show CGS strategy with DQP of 2 and 6, respectively. Curves show that coding efficiency decreases with smaller DQP. The MPEG-4 like FGS coding is represented with orange curve and here only the base layer is used for motion-compensated prediction. The brown curve shows FGS where the coding efficiency is

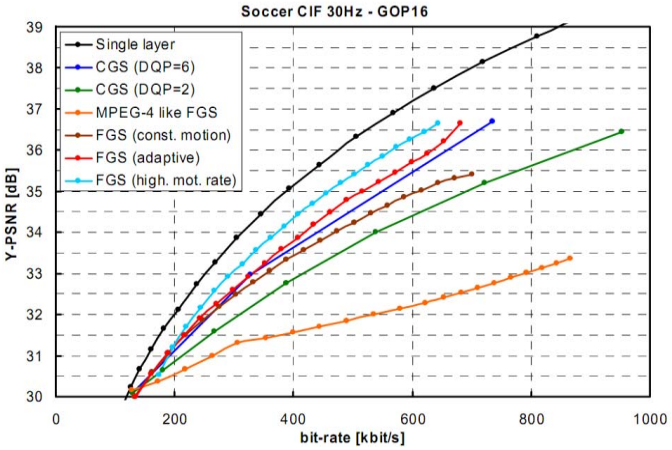


Fig. 9 Comparison of the different SNR scalable coding strategies to single-layer coding for the sequence "Soccer" in CIF resolution and a frame rate of 30 Hz [13]

improved by using references with higher quality for motion-compensated prediction of temporal refinement pictures. The red curve represent improvement of FGS coding with the concept of PR (progressive refinement) slices where the NAL units can be contracted to reduce the bit-rate. The light blue curve shows the adaptive FGS concept which adjusts the ratio between motion and texture data in order to trade-off the coding efficiency for different rates.

IV. RESULTS OF COMBINED TYPES OF SCALABILITY

Temporal, SNR and spatial scalability can be combined and the results for coding efficiency are shown in Fig. 10. The black curve represents coding efficiency for single-layer coding which are compared to the coding efficiency for spatial, SNR and combined scalability for sequence "Soccer" with intra period 1.07 s (64 pictures at 60 Hz). All encodings use dyadic hierarchical prediction structure with a GOP size of 32 pictures.

Single-layer coding supports limited set of points where each represents a separate bit-stream. The red curves represent bit-streams with SNR scalability for QCIF, CIF and 4CIF spatial resolutions. SVC allows rate intervals and for each rate in that interval a part of bit-stream can be extracted. Spatial scalable coding is represented with three blue curves which show three generated bit-streams. The green curves with different spatial-temporal rate points show the coding efficiency of combined scalable coding. The amount of rate points is determined depending on the type of application. There are more substreams that can be extracted from combined scalable bit stream than from single-layer coded streams.

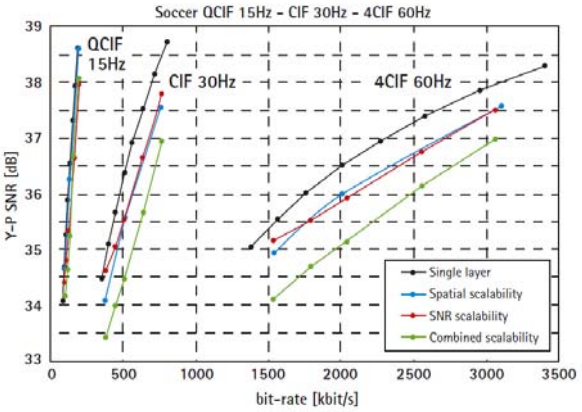


Fig. 10 Performance of combined scalability coding compared to single-layer, spatial scalability and SNR scalability [13]

V. CONCLUSION

Scalable video coding extension of H.264/AVC gives better results in efficiency coding than former standards. The most important improvements are hierarchical prediction structure which enables better coding efficiency than IBBP structure and the concept of key pictures. It is shown that these improvements give 1 dB better results in PSNR gains compared to IBBP coding. Also new methods for spatial coding improve the coding efficiency of spatial and quality scalable coding methods and the results are presented. The residual and motion information can be reused at enhancement layers which gives higher coding efficiency results because of the use of correlations between different layers.

REFERENCES

[1] ITU-T Rec. & ISO/IEC 14496-10 AVC, "Advanced video coding for generic audiovisual services", Version 3, 2005

[2] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, A. Luthra, "Overview of the H.264/AVC Video Coding Standard", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 13, No. 7, July 2003, pp. 560- 576

[3] H. Schwarz, D. Marpe, T. Wiegand, "Overview of Scalable Video Coding Extension of the H.264/AVC Standard", IEEE Transactions on Circuits and Systems for Video Technology, Vol.17, No.9, September 2007, pp.1103-1120

[4] H. Schwarz, T. Hinz, H. Kirchhoffer, D. Marpe, T. Wiegand, "Tehnical Description of the HHI Proposal for SVC CE1", ISO/IEC JTC 1/SC29/WG11, Doc. M10569, March 2004

[5] SVC: Scalable Extension of H.264/AVC, http://www.hhi.fraunhofer.de/en/departments/image-%20processing/image-video-coding/scalable-video-coding/svc-scalable-extension-of-h264avc/ , Accessed: March 2012

[6] N.S. Narkhede, R. Prasad, "Scalable Video Coding and Applications", Proceedings of SPIT-IEEE Colloquium and International Conference, Mumbai, India, 2007-2008, pp.103-109

[7] H. Schwartz, M. Wien, "The Scalable Video Coding Extension of the H.264/AVC Standard", IEEE Signal Procesing Magazine, Vol.25, No.2, March 2008, pp. 135-141

[8] J. Rieckh, "Scalable Video for Peer-to-Peer Streaming", Master's thesis, Institute of Communications and Radio-Frequency Engineering Technical University of Vienna, Summer, 2008

[9] M.A.J. Barzilay, J.R. Taal, R.L. Lagendijk, "Subjective Quality Analysis of Bit Rate Exchange Between Temporal and SNR Scalability in the MPEG4 SVC Extension", IEEE International Conference on Image Processing, 2007. ICIP 2007, San Antonio, USA, pp.II – 285-288

[10] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MCTF", Proc. ICME 2006, July 2006

[11] Y.-H. Chen, T.-D. Chuang, Y.-J. Chen, L.-G. Chen, "Bandwidth-Efficient Encoder Framework for H.264/AVC Scalable Extenision", Ninth IEEE International Symposium on Multimedia, December 2007, pp. 401-406

[12] H.-C. Huang, W.-H. Peng, T. Chiang, H.-M. Hang, "Advances in the Scalable Amendment of H.264/AVC", IEEE Communications Magazine, Vol. 45, No. 1, January 2007, pp. 68-76

[13] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of Scalable H.264/MPEG4-AVC Extension", Proceedings of IEEE International Conference on Image Processing (ICIP '06), Atlanta, USA, October 2006, pp. 161–164