Spatial Enhancement Layer Utilisation for SVC in Base Layer Error Concealment

Mikko Uitto

VTT Technical Research Centre of Finland
Kaitoväylä 1, 90570 Oulu, Finland

mikko.uitto@vtt.fi

Janne Vehkaperä

VTT Technical Research Centre of Finland
Kaitoväylä 1, 90570 Oulu, Finland

janne.vehkapera@vtt.fi

ABSTRACT

The Scalable Video Coding (SVC) has been recently added as an extension to H.264/AVC standard. This extension allows both bit rate and device capability adaptation which are desirable features especially in error-prone wireless heterogeneous networks. This paper investigates the spatio-temporal error concealment techniques for packet losses in wireless IP networks, where scalable video coding (SVC) can be used. Three methods are introduced: pixel-value interpolation, frame copy and a new method, which utilises the correctly received spatial enhancement layer information if the corresponding base layer is missing. Unlike the new method, the traditional methods discard the corresponding enhancement layer data in the case when the base layer is missing. The simulation results indicate that enhancement layer utilisation provides better results in the case of a missing base layer than the traditional error concealment methods improving the image quality on average 2 dB.

General Terms

Algorithms, Standardization.

Keywords

H.264/AVC, SVC, error resilience, adaptive video.

1. INTRODUCTION

Modern technology has made it possible to transfer real-time video to various mobile terminals in wireless networks. The same video can be streamed to low bit rate mobile phones with low quality as well as high bit rate televisions with extremely high quality. Scalable video coding has enabled adaptive video transmission where adaptation to the channel can be done easily without re-encoding. Several types of sub-streams can be decoded from a single encoded stream with the desired spatial,

temporal and quality scalabilities that are suitable for the specific applications.

However, the wireless transmission environment is very error-prone compared to the wired systems. Bit inversions and packet losses can lead eventually to an unacceptable video stream, where parts or even whole frames might be missing. Since real-time demands create their own limitations to transmission systems, re-sending the missing packets is not usually possible especially when broadcasting. As a result, error-resilient tools for the encoder to protect from errors and concealment methods for the decoder are needed in order to achieve sufficient video quality.

The previous SVC standard draft included four different error concealment methods [1], which will be discusses in more detail in subsection 2.2.3. However, these methods had a complex structure and were based on the open-loop structure of the decoder. The structure of the decoder has changed afterwards from open-loop to closed-loop which affects also to the error concealment procedure. The authors of [2] introduced an error concealment technique for SVC, which utilises the spatial enhancement layer information for I and P slices following the open-loop structure of the decoder. Some of the research has focused also on inter-layer correlation in the SVC [3] as well as NALU utilization process [4]. However, many of the error resilience and concealment techniques follow the older open-loop approach. The methods introduced in this paper are developed to the newest JSVM reference codec [5] and their functionality concerns also B-slices, which can be very common element in the GOP format.

This paper is organised in the following way. The second section introduces the scalable video coding extension of H.264/AVC focusing on the spatio-temporal scalabilities and video packetisation. A small survey to existing error concealment techniques is also made. The third section illustrates the implemented error concealment techniques, which are evaluated in the fourth chapter. The fifth and at the same time final section draws a conclusion from the simulation results.

2. SCALABLE VIDEO CODING

The latest video compression standard H.264/AVC has been developed jointly by the ISO/IEC MPEG and ITU-T VCEG standardisation organisations providing improved coding

efficiency by using state-of-the-art coding tools [6]. It provides a network-friendly packet-based video representation supporting both conversational (video telephony) and non-conversational (streaming media, storage) applications. This work focuses on Annex G of the H.264/AVC standard, which defines the scalable video coding extension [7].

The adaptive video transmission methods, such as rate control, transcoding or bitstream switching, offer video adaptation for the channel, but each method has their shortcomings. Scalable video coding (SVC) with three scalability schemes combined or not combined together allows video stream adaptation without the need for the above-mentioned alternative methods. An adaptation to device capability as well as to bit rate can thereby be reached. Various use scenarios exist, such as video broadcast or unicast, video conferencing and surveillance.

Scalable video needs to be encoded only once at highest resolution or with the best quality, but multiple scalable sub-streams can be decoded depending on the target characteristics [8]. With the help of this concept, flexibility and adaptability for several heterogeneous networks can be achieved. Scalable video coding provides many alternatives and possibilities for future video transmission in heterogeneous, error-prone networks – with the price of a heavier encoding process.

2.1 Spatio-temporal scalabilities

Temporal scalability becomes useful in applications where a lower frame rate is needed whereas spatial scalability affects to the video resolution. Rate adaptation can be achieved by using temporal scalability. Naturally, the needed bandwidth increases depending on the number of temporal or spatial stages.

A hierarchical prediction structure can be used with the concept of B pictures in H.264/AVC and it provides excellent coding efficiency both in dyadic and non-dyadic cases [9]. Basically, any picture can be marked as a reference picture and used for motion compensated prediction. Temporal scalability has been supported in MPEG-1, MPEG-2 Video, MPEG-4 Visual, H.262 and H.263 to some extent, but thanks to reference picture memory control H.264/AVC offers more flexibility for temporal scalability. A good example of the flexibility in SVC is that spatial scalable coding supports arbitrary resolution ratios as long as the resolution does not decrease when moving from base layer to enhancement layer.

Multiple layer coding is the key element in SVC. Inter-layer prediction mechanisms are implemented to improve the coding efficiency. The idea is to use lower layer information as much as possible to obtain improvements in the rate-distortion efficiency in the enhancement layers. It is noticeable that temporal predictors can provide a better approximation of the original signal than the up-sampled versions obtained by the base layer reconstruction.

Texture data can be predicted from the up-sampled reference layer texture, where intra-coded macroblocks are used, which makes the process called inter-layer intra prediction. Similar to inter-layer inter prediction, the coding method of the reference layer macroblock defines which prediction method is used.

With the help of the base mode flag, the macroblock of the enhancement layer is predicted from the macroblock of the reference layer and coded depending on the coding structure in the reference macroblock. Consequently, data partitioning and up-sampling is used to the reference layer macroblocks to get the motion vectors for enhancement layer macroblocks. This whole process is called inter-layer motion prediction [9].

Inter-layer residual prediction depends on the value of the residual prediction flag to inform that the residual of the wanted enhancement layer macroblock is predicted from the up-sampled reference layer residual. Only the corresponding difference signal needs to be coded in the enhancement layer [9].

2.2 Bitstream structure

The bitstream in H.264/AVC is separated into VCL and NAL. Generally speaking, all the source data is created as a representation to VCL while NAL formats and encapsulates this data and also generates effective header information, which can be used in several different systems. Since these units are used in transport layer mapping, good flexibility for operation over a variety of networks can be achieved [10]. Figure 1 represents the interconnection between VCL and NAL.
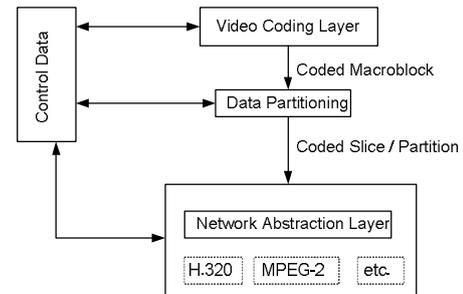

Figure 1. Interconnection between VCL and NAL.

2.2.1 Network abstraction layer

All the coded video information is organised into network abstraction layer units (NALUs), which are the actual packets to be passed to transmission protocols (e.g. RTP) containing an integer number of bytes. Each NALU starts with a prefix code (0x0001), which will not take presence in any other locations in the bitstream. The header byte indicates the type of this payload among many other fields. This allows an effective customisation of the use of VCL in a variety of network systems.

In H.264/AVC one NALU can be divided into one or more RTP packets or many NALUs can be put into one RTP packet [11]. The general NAL unit packet format consists of a header byte and VCL or non-VCL data. The actual coded video data lies in the VCL partition.

2.2.2 Video coding layer

To address the need for flexibility and customisability, the H.264/AVC design covers a VCL, which is designed to efficiently represent the actual video content. The particular part is located after the five-byte NAL unit type identifier. When speaking of the general structure of VCL in H.264/AVC,

it follows the so-called block-based hybrid video coding approach [9]. There are similarities in the VCL design between prior video coding standards such as MPEG-2, but H.264/AVC includes new features in order to exceed the older standards in compression efficiency as well as adaptability and flexibility.

2.2.3 Existing error concealment methods

The scalable extension of the H.264/AVC standard draft earlier included four different schemes for error concealment in the decoder. Two of these methods operate in the intra layer while the other two can overstep the scalability layers and function between layers. These latter two are called inter layer error concealment methods. Overall, the general structure of these methods was very heavy in the codec resulting in a huge number of extra code lines. In addition, the earlier structure for the codec was based on the open-loop principle, which means that each layer was decoded in a separate way and combined finally with the higher layer picture. The structure has changed to the closed-loop principle, which means that the decoding process is done for the whole AU at a time, not layer by layer. The presented draft error concealment methods also had restrictions to their use scenarios (number of enhancement layers) and therefore the reference codec does not include these methods in its releases anymore.

Frame copy (FC) is an intra layer error concealment method that functions both in base and enhancement layers. Its principle is simply to copy the missing pixel values to the erroneous frame from the corresponding pixel in the first frame in the reference picture list 0. Since the motion between frames is not considered at all, this method is not very effective for use in video sequences that contain a lot of motion [1].

Temporal direct motion vector generation (TD) is also an intra layer error concealment method. The missing frame is predicted using two reference picture lists and the desired missing motion vectors will be generated by scaling the motion vectors in the missing frame temporally between these two reference lists [1]. TD will give good results if the motion in the reconstruction area is relatively uniform.

Motion and residual up-sampling (BLSkip) is an inter layer error concealment method and its purpose is to conceal a lost spatial enhancement layer from the predicted P- or B-pictures. The residuals and motion vectors of the base layer will be up-sampled to higher resolution for the enhancement layer [1].

The last method, reconstruction base layer up-sampling (RU) is an inter layer error concealment method, where the base layer picture is reconstructed and up-sampled using a 6-tap H.264/AVC filter for the lost enhancement layer picture [1].

3. IMPLEMENTED ERROR CONCEALMENT SCHEMES

All the implementations were added to Joint Scalable Video Model (JSVM) reference codec version 9.15. This is the reference software for the scalable video coding project from Joint Video Team (JVT) working groups [5].

3.1 Traditional error concealment techniques

All the implementations are based on a map where all the macroblocks are marked as

$$0 = correctly\ received\ MB$$
$$1 = lost\ MB$$
$$2 = reconstructed\ MB$$

With the help of this map, all the error concealment techniques can be performed only for the missing macroblocks. The map is actually a two-dimensional array, where the layer identification as well as the location of a missing macroblock travels together. The array is initialised at the beginning of each access unit.

3.1.1 Pixel-value interpolation

In order to implement an error resilient decoder, a concealment algorithm for I frames was needed. The problem in using a slice copy to I-frames is that there are no available frames in the reference list. One solution is to use the mentioned pixel-value interpolation for the I-frames. Figure 2 shows an example of the interpolation.
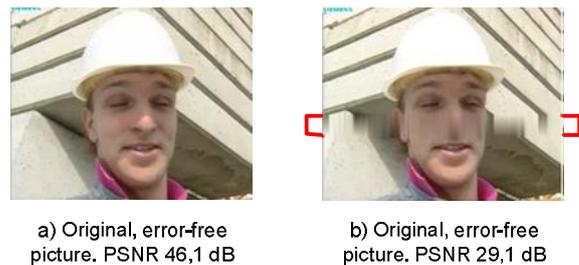

a) Original, error-free picture. PSNR 46,1 dB   b) Original, error-free picture. PSNR 29,1 dB

Figure 2. An example of pixel-value interpolation.

The interpolation function uses the correctly received and decoded macroblock and pixel areas from the same frame where the missing slice is located. Above and below macroblocks are used as sources to conceal the intermediate macroblock. If only the other source is available, the interpolation becomes stretched. The worst case is when both sources are unavailable, which means that macroblock lines are skipped as long as the correctly received macroblock is found. The reconstruction is done in raster scan order. The final thing to do in the interpolation function is to mark the concealed macroblocks as reconstructed in the missing macroblock map.

3.1.2 Frame copy

Frame copy is a very convenient solution for replacing the missing picture area in the case of packet loss. The missing slice is copied from a reference list and the results are excellent if there is no motion inside the GOP. Naturally, when the GOP size is long and there is plenty of motion, Frame copy gives poor results. The reference list picture can be located either temporally earlier or later in the video stream as will be later pointed out. Figure 3 represents an example of frame copy.

a) Original, error-free
picture. PSNR 42,4 dB.

a) Original, error-free
picture. PSNR 25,7 dB.

Figure 3. An example of frame copy.

There are two reference frame lists available, which are defined in the standard and can be applied to the frame copy. The first list, reference list 0 is used for P or B pictures whereas reference list 1 can be utilised only for B-pictures. The difference between these two lists is that list 0 uses the temporally earlier key pictures (I or P) in the GOP and the reference picture list 1 uses the temporally nearer reference pictures, which can also be a B-picture [6]. Using reference list 1 will give smoother results since the frame to be copied is not as far away from the picture to be reconstructed. Figure 4 represents the algorithm logic of the traditional methods.

3.2 Enhanced error concealment method

It is commonly assumed that the missing base layer slice leads to rejecting also the dependent enhancement layer slice. In many cases, this is true if the higher layer needs the base layer data for the picture formation. The macroblocks in a frame can be coded differently, as was presented earlier. The basic assumption is that the blocks including lots of motion are intra-coded while the static areas are inter-coded, which are predicted from other pictures from the same layer and also from the lower layer, since the hierarchical prediction structure is usually used in SVC. For example, when looking at the whole *Foreman* sequence at a frame level, it is noticeable that the reference pictures have many intra-coded blocks.
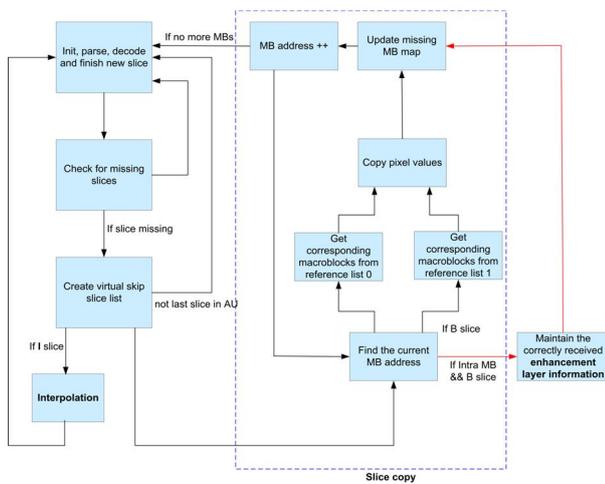

Figure 4. An algorithm representing the SVC decoder with the implemented error concealment features.

The simplified block diagram of the spatial enhancement layer utilisation process can be seen in Figure 4. This particular method can be seen in the far right. In general, the enhancement layer utilisation process is used when the picture type is B and current macroblock is intra-coded. The majority of the frames in the video sequence are usually B-pictures with hierarchical prediction structures, which makes this method rational to use.

Basically enhancement layer utilisation means that the texture elements of the correctly received enhancement layer data are gathered and used in the reconstruction. Since some of the base layer data is still needed, the macroblock copy is applied to all the blocks that are inter-coded.

Figure 5 shows an illustrative example of the enhancement layer utilisation. As can be seen, the visual results are very good. The particular image is a reference picture and it has lots of intra-coded macroblocks especially in the face area since it contains motion from one frame to another. The mouth area is nicely reconstructed and the blocking phenomenon can be seen well only in the violet collar of *Foreman*.
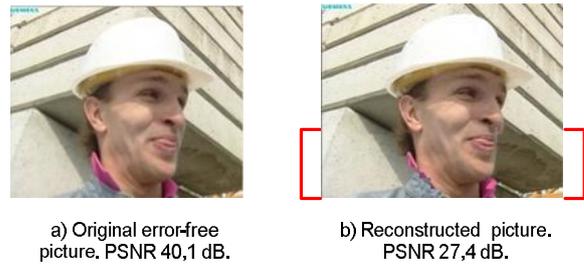

a) Original error-free
picture. PSNR 40,1 dB.

b) Reconstructed picture.
PSNR 27,4 dB.

Figure 5. An example of enhancement layer utilisation.

4. SIMULATION RESULTS

Three different video clips that were used in the simulations: *Foreman*, *Soccer* and *Hall*. These three were selected for the following reasons. *Foreman* is a widely used test video sequence among developers of video codecs since it contains both small and relatively large motion inside the clip of 300 frames. The *Soccer* sequence differs from the *Foreman* especially in the amount of motion and contains more moving elements. In contrast to the first two sequences, the *Hall* test stream has a static camera with a couple of moving objects. Simulation parameters are shown in Table I.

The test stream was 300 frames long (10 seconds) for all the three test sequences regardless of the choices for coding parameters. Naturally, all the encoded versions were made with a compliant JSVM 9.15 encoder from the same reference software as the decoder. The packet droppings were centralised and restricted only to the base layer. Packet loss ratios from two to ten with the interval of one were performed in all the test cases.

Table 1. Simulation parameters

Input video	*Foreman*	*Soccer*	*Hall*
Spatial layers	2	2	2
Number of frames	300	300	300
BL resolution	QCIF	QCIF/CIF	QCIF
EL resolution	CIF	CIF/4CIF	CIF
Slices in BL	3	3	3
Slices in EL	9	9	9
Frame rate BL	30	30	30
Frame rate EL	30	30	30
GOP size	8/16	8/16	8/16
QP BL	38	38	38
QP EL	20	20	20
PLR BL	0…10	0…10	0…10
PLR EL	0	0	0
Bit rate BL(kbit/s)	420	350	320
Bit rate EL(kbit/s)	2000	2100	1880

4.1 Evaluation of interpolation and frame copy

Figure 6 illustrates the visual results of the implemented methods. As can be seen from Figure 6 d) the spatial enhancement layer utilisation provides the best results both from PSNR as well as visual aspects.

Figure 7 shows the simulation results from the interpolation function, which is designed to function for the missing slices and macroblocks for the I-slices. In order to reach a stable decoder in the simulation environment when performing the other test cases, it is rational that the packet losses are centralised to the whole base layer and not only for P and B-slices. As can be seen, the *Soccer* sequence provides the best results for the interpolation, beating the competitors by 3-4 dBs. The main reason for the superiority of *Soccer* is that it contains lots of motion both from the camera as well as the contents of the video with moving objects and changing textures in the surroundings.

Figure 7 shows the similar result from frame copy. Again, in order to reach a better replicate from an actual real-time transmission environment, the interpolation method was used for I-pictures in case any missing slices exist for these types of pictures. Similarly as in the interpolation, all the diagrams decrease quite linearly as the packet loss ratio increases. This time, the *Hall* sequence provides the best results and works very well for this video sequence. Even the 10% loss ratio will not drop the average PSNR under 30 dB, under the limit of acceptable quality.

4.2 Evaluation of spatial enhancement layer utilisation

Figure 8 represents the simulation results from the spatial enhancement layer utilisation process. Since macroblock copy is performed to all P-slices as well as inter-coded blocks in B-slices, the *Hall* sequence provides again the best results. All the average PSNR values for all three sequences remain over 30 dB at the 8% packet loss rate, which was classified as the lowest limit for good video quality.

The effect of using higher resolution both in the base layer (CIF) as well as in the enhancement layer (4CIF) can be seen in Figure 9, which shows in (a) the difference when using GOP sizes 8 and 16, which is approximately 1 dB. Furthermore, as can be seen from (b) a higher base layer has approximately a 1-2 dB improvement for the average PSNR compared to the traditional QCIF-CIF category.

Figure 10 shows the simulation results both with GOP sizes 8 and 16 between the three implemented methods. As can be seen, the GOP size has approximately a 0.5 dB impact on the results concerning especially the *Foreman* sequence. Interpolation gives the worst results while enhancement layer utilisation is the best technique. The difference between interpolation and the utilisation process is up to 4 dB, which is quite significant. Comparison between the utilisation and slice copy brings a 1-2 dB improvement, which is quite rational since the enhancement layer utilisation uses slice copy to inter-coded macroblocks.

a) Original, error-free picture.　　b) Interpolation.　　c) Frame copy.　　d) Spatial EL utilization.

Figure 6. Reconstructed PSNR values: a) 38.2 dB, b) 21.5 dB, c) 23.0 dB and d) 26.6 dB.
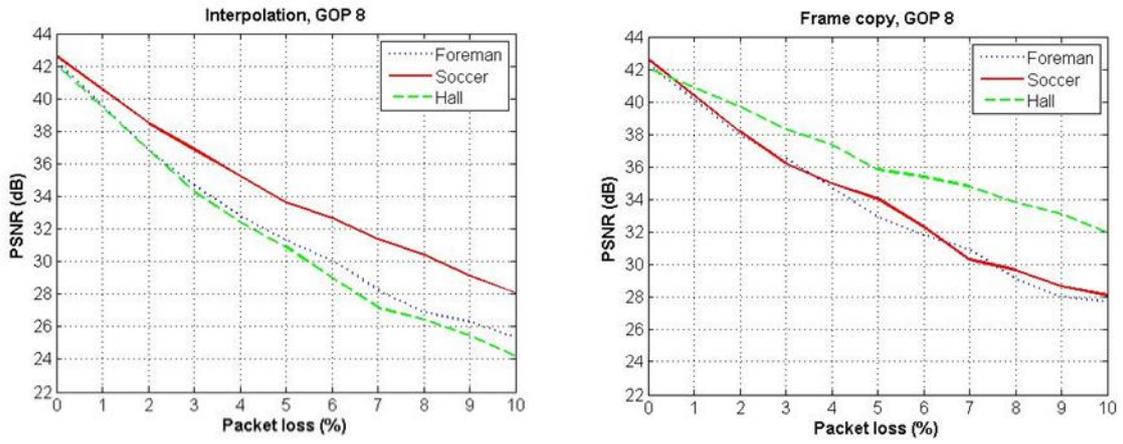
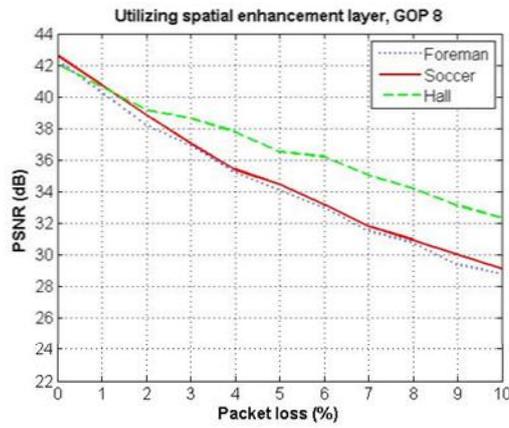**Figure 7. Simulation results of the traditional methods with GOP size 8.**

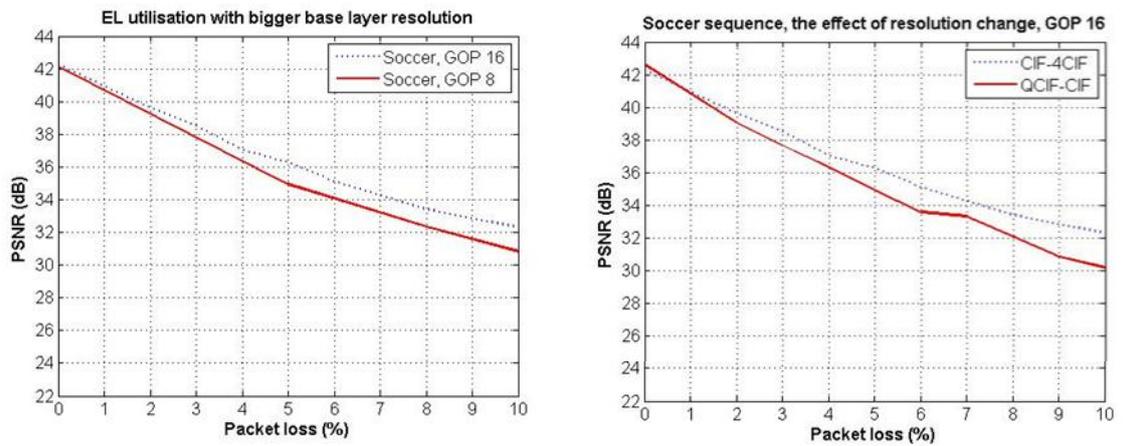**Figure 8. Simulation results of spatial enhancement layer utilisation using GOP size 8.**

Figure 9. Simulation results of the enhancement layer utilisation with larger resolutions.
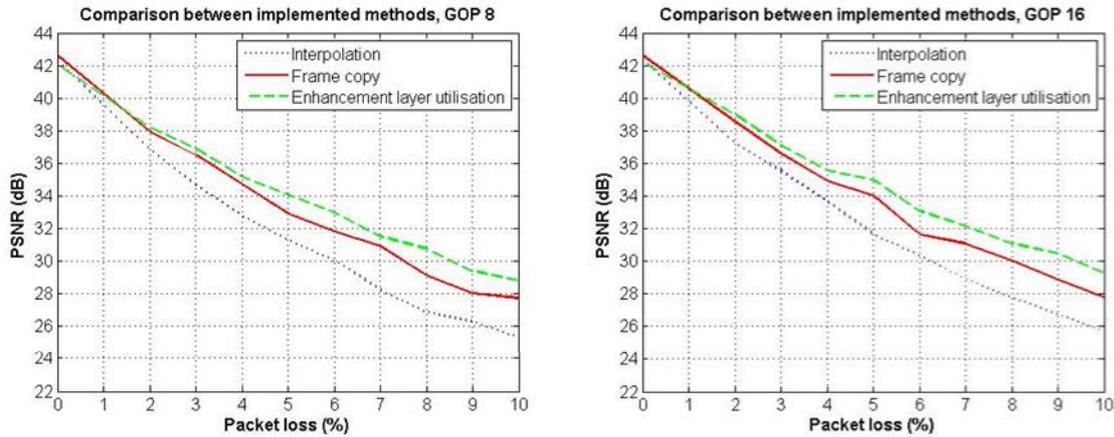
Figure 10. Comparison between the implemented methods with *Foreman*.

5. CONCLUSION

This paper studied scalable video coding in wireless packet-switched networks focusing on spatio-temporal error concealment techniques. The paper introduced three error concealment methods: pixel-value interpolation, frame copy and spatial enhancement layer utilisation, which all were implemented to the JSVM 9.15 reference codec. Frame copy utilises temporally earlier or later picture data whereas interpolation and spatial enhancement layer utilisation uses the available picture data from the current frame. All the packet losses in the simulations were restricted to the base layer.

All the tests were performed with both GOP sizes 8 and 16. The simulation results indicate that the GOP length does not have a particularly large effect on the results. The spatial enhancement layer utilisation provides the best video quality compared to the traditional methods. The average difference between the enhancement layer utilisation and frame copy was 2 dB. The results are better also from a visual viewpoint and the comparison with individual pictures can improve the image quality even 4 dB.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] Chen Y., Xie K., Zhang F., Pandit P. & Boyce J. (2006) Frame loss error concealment for SVC. Journal of Zhejiang University – Science A 7, p. 677-683.

[2] Keränen T., Vehkaperä J. & Peltola J. (2008) Error Concealment for SVC Utilizing Spatial Enhancement Information. In 4th International Mobile Multimedia Communications Conference, MobiMedia 2008, Oulu, Finland.

[3] Park C. S., Wang T. S. & Ko S.J. (2007) Error Concealment Using Inter-layer Correlation for Scalable Video Coding. ETRI Journal, Vol. 29, p. 390-392.

[4] Ngyen D. T., Shaltev M. & Ostermann J. (2006) Error Concealment in the Network Abstraction Layer for the Scalability Extension of H.264/AVC. In: ICCE '06, First International Conference on Communications and Electronics, October 10-11, Hanoi, p. 274-278, Vietnam.

[5] SVC Reference Software (JSVM software). [Online]. Available:http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm

[6] Sullivan G. J., Topiwala P. & Luthra A. (2004) The H.264/AVC Advanced Video Coding Standard: Overview and Introduction to the Fidelity Range Extensions. In: SPIE Conference on Applications of Digital Image Processing XXVII, August 2-4, Denver, p. 454-474, USA.

[7] ITU-T (2007) Series H: Audiovisual and Multimedia Systems, Recommendation H.264: Advanced video coding for generic audiovisual services, 564 p.

[8] Ohm J.-R. (2005) Advances in Scalable Video Coding. Proceedings of the IEEE, Vol. 93, p. 42-56.

[9] Schwartz H., Marpe D. & Wiegand T. (2007) Overview of the Scalable Video Coding Extension of the H.264/AVC Standard. IEEE Transaction on Circuits and Systems for Video Technology, Vol. 17, p. 1103-1120.

[10] Ueda S., Shigeno H. & Okada K. (2007) NAL Level Stream Authentication for H.264/AVC. IPSJ Digital Courier, Vol. 3, p.55-63.

[11] Wenger S., Wang Y.-K. & Schierl T. (2007) Transport and Signaling of SVC in IP Networks. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 17, p. 1164-1173.