Error Concealment for SVC Utilizing Spatial Enhancement Information

Tommi Keränen Janne Vehkaperä Johannes Peltola

VTT Technical Research Centre of Finland
Kaitoväylä 1, FI-90571Oulu, Finland

tommi.keranen@vtt.fi janne.vehkapera@vtt.fi johannes.peltola@vtt.fi

ABSTRACT

In this paper packet loss error concealment for video sequences compressed using spatial scalability is investigated. Slice support is implemented into the JSVM reference codec of the scalable extension to H.264/AVC video coding standard. The non-normative error concealment scheme introduced in the codec is developed further, adding to it the capability to also consider correctly received slice information from the same frame to conceal lost frame areas. In case of lost base layer slices further improvement on the reconstruction is then achieved by using the correctly received spatial enhancement information for the same frame. The proposed enhancements focus on packet loss concealment on the base layer of I and P-coded frames, where the greatest improvements to the original scheme were identified. Simulation results for given packet loss model indicate on average 2dB improvement over original scheme in the target error scenario.

General Terms

Algorithms, Standardization, Theory.

Keywords

H.264, SVC, adaptive multimedia, error concealment.

1. INTRODUCTION

Streaming video content over IP networks is a challenging task, requiring not only a bitstream format suitable for packet switched networks, but also tools to cope with the highly heterogeneous nature of environments like the wireless Internet, where available bandwidth, error probabilities and error patterns might change drastically in time and between network nodes. In addition to these problems the requirements and limitations of each end-user can be just as manifold, whether it is the desired resolution, bitrate, image quality, and complexity of the compression or real-time restraints.

Scalable video coding techniques solve many of these problems, allowing multiple different video representations to be extracted and decoded from a single bitstream. A scalable extension to H.264/AVC video coding standard is being developed by Joint Video Team of ISO/IEC MPEG and ITU-T VCEG. This extension brings several scalability technologies into a single framework, which provides several tools to achieve high levels of scalability while retaining the high compression performance of H.264/AVC [1].

The cost of these new abilities in the scalable video codec (SVC) is the amount of extra complexity that the scalable structure introduces into the compressed video, mostly in form of extra prediction layers in the already complex prediction chains between frames and picture elements present in H.264/AVC. Many design choices in the standard's syntax reflect the need to add extra robustness and error resiliency to battle the vulnerabilities of a compressed bitstream against random bit errors and packet losses, but even with all that, there is still a need for concealment methods that can conceal errors that can't be otherwise dealt with.

In the second section of this paper a short overview of SVC is provided in the scope of this paper. The third section goes into the modifications done to the Joint Scalable Video Model (JSVM) reference codec and in fourth part simulations are performed. Fifth section draws conclusions from the test results.

2. SCALABLE VIDEO CODING

2.1 Spatial Scalability

This paper concentrates around the notion of spatial scalability layers. The video source is compressed so that representations of several different resolutions can be extracted and decoded from the coded bitstream. The lower resolution base layers can be used to predict the higher resolution layers, thus increasing the coding efficiency of these enhancement layers.

There are three ways to perform the inter-layer prediction, and the encoder can make a choice between these methods on macroblock-to-macroblock basis, depending on which results in the most suitable rate-distortion ratio. First of the methods uses the upsampled decoded texture elements of the base layer as the prediction for the enhancement layer. Second method uses the prediction residual of the base layer for the same task, and the third method uses the upsampled motion information. All of these predictors must be upsampled to fit the higher-resolution layer before use, and so their accuracy for high-frequency elements is poorer. With motion prediction arbitrary motion refinement bits can be added to the bitstream to fine-tune the inter-layer prediction before decoding.

Mobimedia 2008 July 7-9, 2008, Oulu, Finland.
 ISBN 978-963-9799-25-7/08/07
DOI 10.4108/ICST.MOBIMEDIA2008. 3989

When developing error concealment methods for scalable coding systems, the fact that all these prediction methods can be used side by side, has to be taken into account, as this situation also introduces multiple new paths for error propagation and manifestation. A seemingly high quality reconstruction of the base layer can very well look far from perfect on higher spatial layers.

2.2 Single-loop vs. Multiple-loop Decoding

When streaming video, for example, to a mobile device, the computational power at the disposal of the video decoder is very limited. Because of this it is imperative to make the decoding process as straight-forward and simple as possible. In scalable video coding, when decoding higher spatial layers, it is often necessary to decode the lower spatial layers first to obtain predictors for the higher layers, thus performing the decoding-cycle multiple times for each frame location. In single-loop decoding this is avoided by allowing the inter-layer prediction only for intra coded base layer macroblocks. The decoding complexity is thus lowered to the level of a normal, single-layer video sequence, sacrificing some of the coding efficiency [2].

SVC standard emphasizes the single-loop approach, but the multiple-loop decoding, which doesn't place any restrictions on inter-layer prediction, is still included and is a part of several profiles, as it enables useful functionality that couldn't be realized as easily using the single-loop approach, such as ROI-coding [3]. In the work presented in this paper, multiple-loop coding is used on all examples and simulations.

2.3 Network Abstraction Layer

H.264/AVC standard introduced a new layer of processing next to the video coding layer (VCL). Network abstraction layer (NAL) abstracts the compressed bitstream generated at the VCL, providing a network independent interface between the video coding system and different packet switched networks [4]. This abstraction is reached by wrapping each coded video slice or a set of parameters into a separate package called NAL unit.

Each NAL unit contains a header describing the contents of the packet and a data field containing the actual compressed video information. Contents of a NAL unit can be decoded independently from other NAL units, providing a natural way of partitioning the data in many pieces to among many achieved qualities to increase error resiliency. This functionality is used as basis for the error concealment methods tested in this paper.

2.4 Error Concealment Scheme Introduced in the Scalable Extension

The SVC standard introduces four non-normative error concealment methods to tackle the problem of reconstructing frames lost due packet losses [5]. These methods can be tagged as intra and inter-layer methods depending on whether they use reference frame or base layer information to conceal lost frames on a given scalability layer.

Simplest of the methods is an intra-layer frame copy method (FC), where the first reference frame from the list of temporally preceding frames is simply copied to replace the lost frame. This method, like the developers conclude, produces the weakest overall quality of the concealment from the four methods introduced in the standard.

Another intra-layer method, temporal direct motion vector generation (TD), works by copying the motion vectors from a temporally latter reference frame and scaling them down according to the position of the lost frame, so that the new set of motion information describes the lost frame in case of completely linear movement within the group of pictures. This method can yield good results especially in scenes of low motion, but if the temporal distance to the closest available reference frames is great and the motion in the scene is high and complex, the quality of the reconstruction drops fast.

The two inter-layer methods introduced in the standard work by using correctly received base-layer information to reconstruct lost higher spatial layers. First of these methods, motion and residual upsampling (BLSkip), uses the motion vector and residual information from the base layer, upsampling them to fit the lost spatial layer and then decoding it normally. The second method, reconstruction base layer upsampling (RU), works by decoding the base layer and then upsampling the texture elements directly instead of the motion and residual information.

Generally both of the inter-layer methods yield excellent subjective and objective reconstruction quality even, when high percentage of the spatial enhancement layers has been lost. This is because the reconstructed frame is content-wise an accurate, though somewhat blurry, representation of the lost frame, and momentary degradation of the image precision doesn't register so easily as annoying for the viewer, especially in high-motion scenes. However, as the error percentage in the base layer grows, the efficiency of both of these methods drops quickly, because before any upsampling can take place, the base layer needs to be reconstructed using some of the intra-layer methods. In this case using the TD-method can produce a reconstruction of the target enhancement layer directly and the base layer is discarded as well.

When looking for ways to improve on this scheme, some points of interest can be identified. First of all, none of the methods are able to work their way around lost base layers in I-coded frames, as no reference frame information is available. Secondly, the use of the trivial frame copy method for P-coded frames (potentially over long temporal distances) can lead to very inaccurate reconstruction in high-motion scenes. This also increases the risk of performing error concealment over scene changes in the video sequence. Thirdly, none of these methods consider spatial redundancy as a source for the reconstruction, that is, correctly received information from the same frame.

3. ENHANCED ERROR CONCEALMENT SCHEME

So far the SVC reference implementation hasn't had a functional slice-support, which explains the lack of error concealment methods that would use it to their advantage. In a practical video-coding system meant for streaming applications it is not reasonable to expect that – especially in higher resolution/image quality cases – each scalability layer of a frame would be wrapped into a single NAL packet. To investigate the effects of error concealment on slices using slice information in scalable

environment this situation was remedied by modifying the JSVM reference codec to include slice support for interleaved slices, where slices are arranged as even and odd rows of macroblocks as illustrated in Fig. 1. This way, whenever a frame is divided into several slices, there are always macroblocks belonging to other slices in the neighborhood of each macroblock of the frame, making concealment of random packet losses easier.

Using information from the correctly received parts of the same frame for error concealment is a common technique. In the H.264/AVC standard two non-normative decoder error concealment methods are described, which use spatially available information for the reconstruction [6,7]. These straight-forward methods were also implemented into SVC in order to compare the differences between using spatial and temporal information for error concealment in scalable environment, and to see, if the correctly received spatial enhancement information could be used to improve the concealment of the base layer.

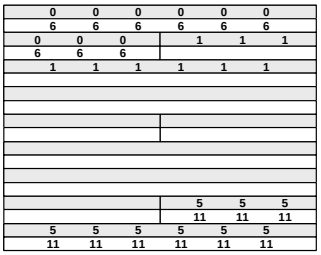
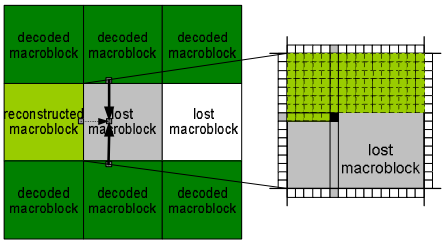


Figure 1. Example of dividing the frame into interleaved slices.

First of the methods, usually used in intra-coded frame areas, is the pixel-value interpolation. Fig. 2 illustrates, how the surrounding correctly received (or previously concealed) image information is used to interpolate a soft gradient over the lost image area pixel by pixel. Previously reconstructed macroblocks are used as a source as well, if there are less than two correctly received macroblocks in the neighborhood.



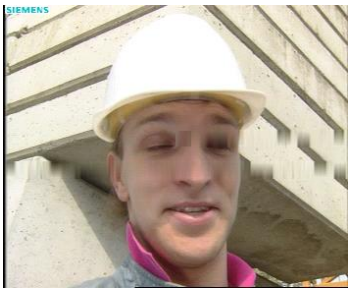


Figure 2. Pixel-value interpolation method.

The second method is used for the reconstruction of lost inter-coded slices. Motion information from each decoded neighbor macroblock is taken in turn and is used to reconstruct the lost macroblock, after which the distortion at the block borders is calculated. The set of motion information leading to smallest distortion is selected for the final reconstruction. Fig. 3 illustrates the process.

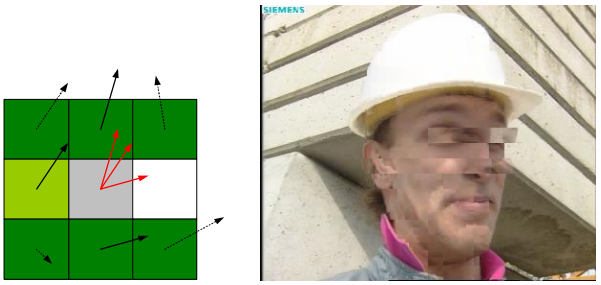


Figure 3. Motion vector evaluation method.

The methods were implemented into the JSVM reference codec as alternatives for the frame copy method, combining them all into an adaptive concealment tool. The method to reconstruct each macroblock on the base layer should be chosen depending on the coding mode of the macroblocks around the lost macroblock and the motion activity in the region. In Fig. 4 the logic that makes the choice between the different methods is illustrated. The motion threshold can be defined to be more or less sensitive to motion, for example depending on the resolution of the video. Threshold length for motion vectors used in these tests was 4 pixels.

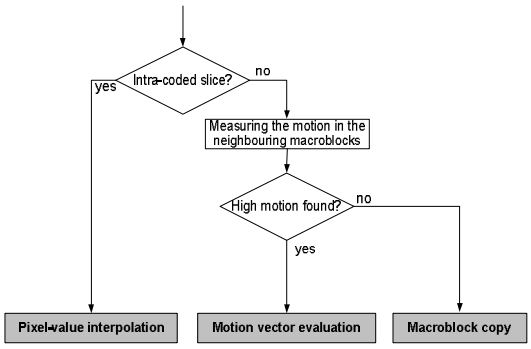


Figure 4. Logic to choose between different reconstruction methods.

4. SIMULATIONS

Packet losses in base layer slices of key frames were identified as the weakest link in the error concealment scheme currently implemented in the JSVM reference codec, and so the performed simulations were concentrated on them. These slices are located in the beginning of long and complex prediction chains and any error in them will greatly affect the quality of a long portion of the video sequence following them, making their reconstruction ability critical for streaming applications.

Figure 5. Reconstruction PSNR-values: a) 38.8 dB, b) 31.5 dB, c) 32.1 dB, d) 32.3 dB, e) 32.7 dB

4.1 Reconstructing key frames

Fig. 4 illustrates the problem of using only reference frame information for the reconstruction of key frame slices. Hall-sequence was coded using GOP-size of 16 frames, and as the feet of the poor fellow in the video are lost due to packet losses, reconstruction is necessary.

In Fig. 5 b) this is done using the frame copy method. Unfortunately in the previous key frame 16 frames earlier the man hasn't appeared on the picture yet save for part of his shoe.

In Fig. 5 c) frame copy is used again, but this time the high frequency information from the correctly received spatial enhancement slice is used as well, leading to a ghost image of the man's feet.

In Fig 5 d) the initial reconstruction is done using the pixel-value interpolation method for the man's feet and macroblock copy to other regions of the slice. Together with the spatial enhancement information a dark blur with some edge information is produced. This situation is very difficult to deal with, as the lost slice contained new picture information that wasn't present anywhere in the previous key frame, but using the pixel-value interpolation method together with the spatial enhancement layer information generally allows for faster recovery and less artifacts in the following frames.

In Fig. 5 e) the image quality of the spatial enhancement layer is higher than that of the base layer, and so more spatial details are reconstructed. In the simulations the improvement affected mostly the initial concealment result, but didn't have a large impact on the over-all time it took to recover from the packet loss. This was mostly due to an issue of texture-leaking, which from time to time appeared in intra-coded regions, when using higher-quality spatial enhancement. This problem is also visible in the picture right below the man's foot. This undermined the effect of the otherwise improved reconstruction quality that can clearly be seen in Fig. 6.

4.2 Results

Several versions of each test sequence were encoded, their essential parameters shown in the table 1. The compressed sequences were then tested by introducing evenly distributed random packet losses to the NAL units containing base layer information of P-coded GOP key frame slices and leaving the enhancement layer information and B-coded slices untouched. Errors were concealed with each set of methods in turn and the tests were repeated for each test sequence and error concealment method 100 times. In the final results averages of the test results over entire sequences are shown.

The first test was performed using the original error concealment methods. In this scheme, the only available method for the target slices was frame copy, in the context of these simulations performed directly to the highest spatial layer. The reconstruction quality achieved by these tests was used as a point of comparison for the other test results. In the second test frame copy was used once again, but this time the correctly received spatial enhancement information was decoded on top of it.

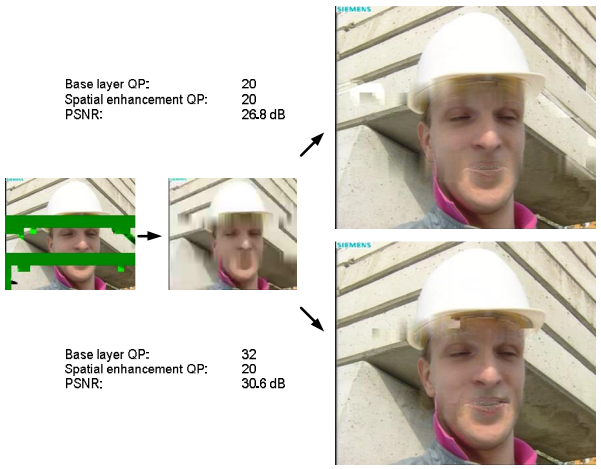


Figure 6. Regaining the spatial details from the spatial enhancement layer after losing a quarter of the base layer in an I-coded frame.

In the third test lost slices were concealed by using the pixel-value interpolation method and in the fourth test this result was further manipulated by decoding the spatial enhancement layer information on top of the gradient.

In the fifth and final test the motion vector evaluation and frame copy methods were used together with the pixel-value interpolation method. The spatial enhancement layer information was added to all concealed information.

In all of the simulations, regardless of the enhancement layer image quality, using the spatial enhancement layer information to improve on the initial base layer reconstruction enabled the biggest improvements to the final image quality of all the available methods.

Comparing the base layer reconstruction methods, a good example of a situation, where the frame copy method produces superior results compared to the pixel-value interpolation method, can be seen in Fig. 7 and Fig. 8. In Hall-sequence the motionless background covers most of the scene, and it can be directly copied from earlier frames for an excellent reconstruction. Using methods such as the pixel-value interpolation on those image areas automatically leads to inferior results even when using the spatial enhancement layer information, as can be seen from the diagrams. In this set of tests, whether the frame copy method was used by itself or in combination with the other implemented methods, didn't affect the results significantly. Using high quality spatial enhancement layer improved the final results, when testing with the longer GOP-size, but with GOP-size of 8 no apparent differences can be seen compared to the spatial layers with similar quantization parameter.

In the second series of tests Foreman-sequence was used. This sequence is shot using a hand-held camera, leading to slight but constant movement of the background, and a more drastic change towards the end. The sequence has also a lot of varying motion in it, making it a challenging target for the concealment methods. Using spatial enhancement information on top of the reconstructed base layer became the largest contributor to the final quality. As the sequence doesn't really have any stationary elements, using frame copy didn't have the same advantage it had

in the previous tests, and as can be seen from Fig. 9 and Fig. 10, the choice of base layer reconstruction method wasn't so crucial this time around. The motion vector evaluation method gave the tests with combined methods some additional edge compared to the other schemes, but especially with the high quality spatial enhancement layers the enhancement information dictated the final reconstruction quality.

Table 1. Coding parameters for the test sequences

Test number	1/3	2/4	5/7	6/8
Test sequence	Foreman	Foreman	Hall	Hall
Spatial layers	2	2	2	2
Base layer QP	20	32	20	32
Base layer resolution	176×144	176×144	176×144	176×144
Slices/frame on base layer	9	9	9	9
Spatial enhancement QP	20	20	20	20
Spatial enhancement resolution	352×288	352×288	352×288	352×288
Slices/frame on spatial enhancement	1	1	1	1
Inter-layer prediction	yes	yes	yes	yes
GOP-size	8/16	8/16	8/16	8/16

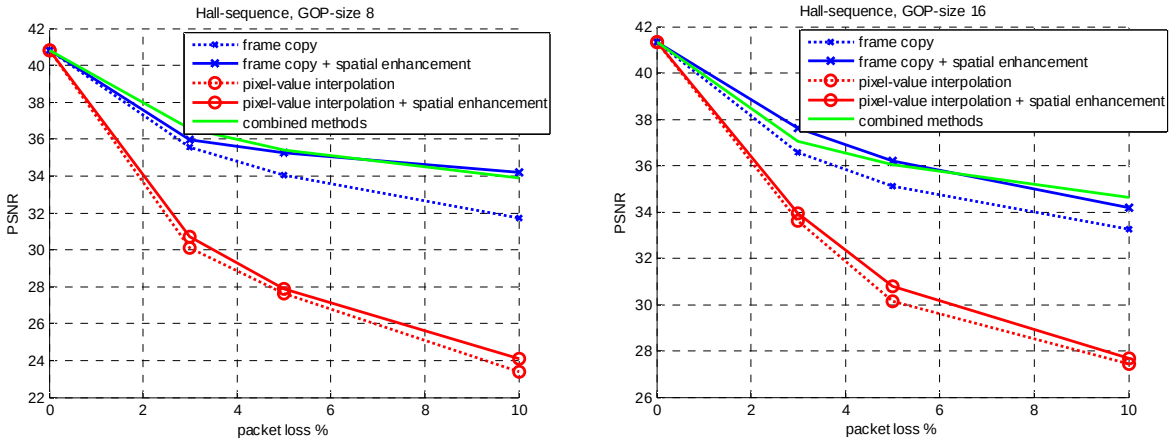


Fig. 7. Tests 1 and 2, *Hall*-sequence with equal image quality on both spatial layers.

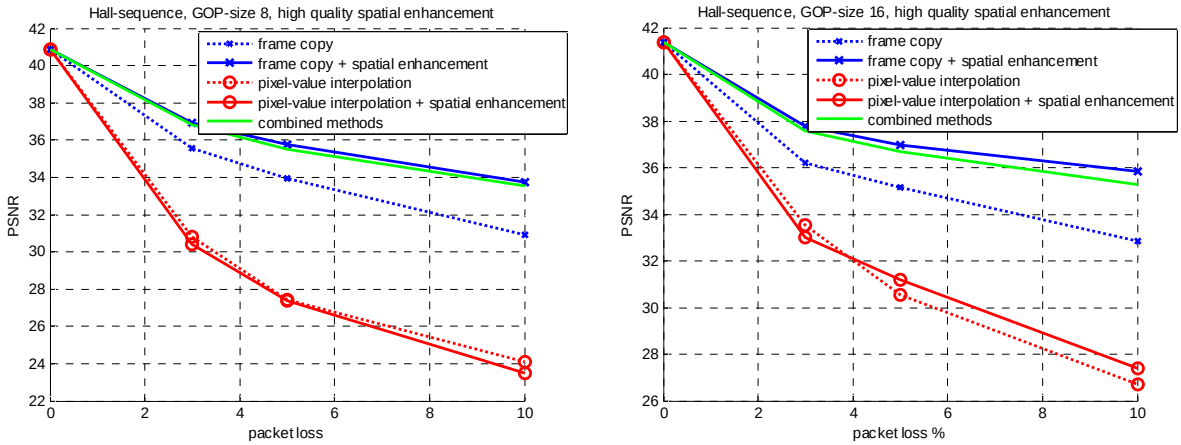


Fig. 8. Tests 3 and 4, *Hall*-sequence with higher quality spatial enhancement layer.

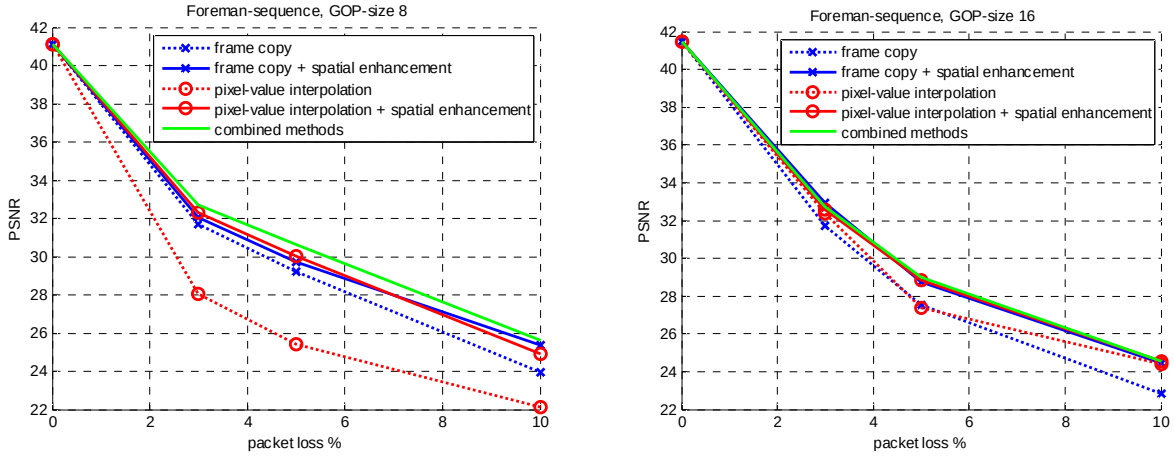


Fig. 9. Tests 5 and 6, *Foreman*-sequence with equal image quality on both spatial layers.

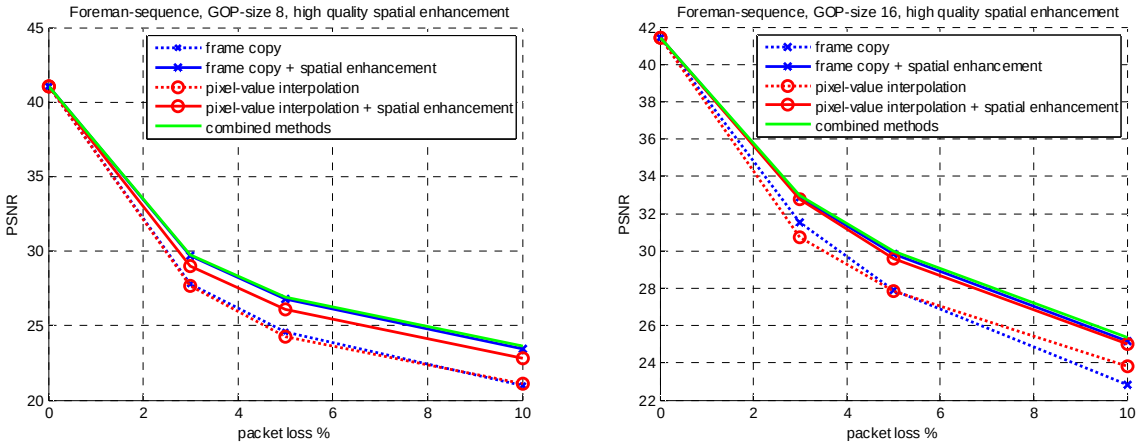


Fig. 10. Tests 7 and 8, *Foreman*-sequence with higher quality spatial enhancement layer.

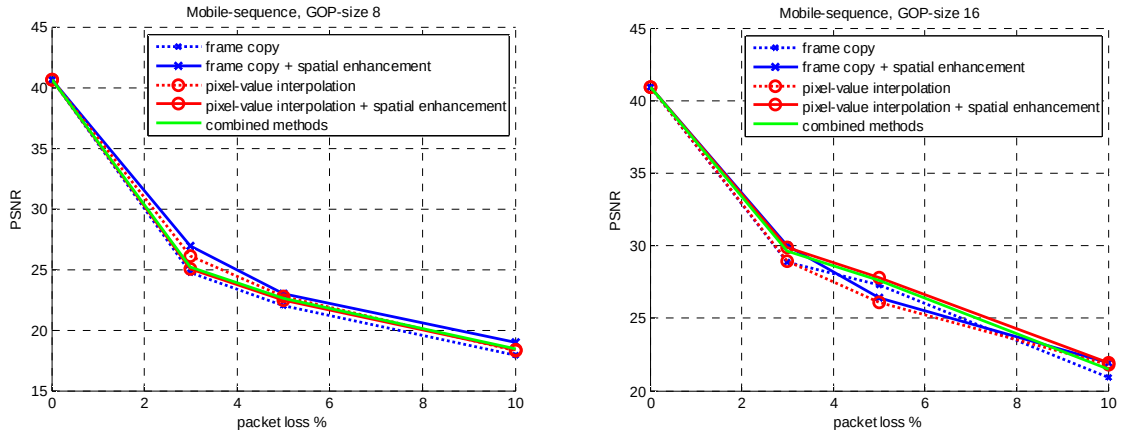


Fig. 11. Tests 9 and 10, *Mobile*-sequence with equal image quality on both spatial layers.

5. CONCLUSIONS

In this paper error concealment in scalable video codec was investigated and simulation results for different error concealment schemes were presented. The non-normative error concealment scheme introduced in the JSVM reference codec was developed further to use not only reference frame information, but also correctly received information from the same frame and from higher spatial layers to deal with packet losses on the base layer of the compressed stream. To make this possible a slice support was also implemented to the JSVM reference codec.

The simulations were focused on packet losses that occur on the base layer of the I and P-coded key frames of groups of pictures. The results show that using the extended concealment scheme to deal with these packet losses improves the reconstruction quality of the sequence on average by 2 dB, when the higher spatial layers are targeted for the final representation.

Most of the quality improvement was due to the use of spatial enhancement layer information. Instead of discarding the enhancement layer information after suffering losses on the base layer, the higher spatial layers were used to recover additional details on top of the base layer reconstruction. In cases, where the image quality of the spatial enhancement layer was higher than that of the base layer, the reconstruction quality could in some cases be improved as much as 5 dB over the frame copy method.

6. REFERENCES

[1] Joint Video Team. 2005. Joint Scalable Video Model (JSVM) 4.0 – Annex G. ISO/IEC MPEG N7556, Oct. 2005.

[2] Schwarz, H., Hinz, T., Marpe, D., and Wiegand, T. 2005. Constrained Inter-Layer Prediction for Single-Loop Decoding in Spatial Scalability. Proceedings of IEEE International Conference on Image Processing (ICIP'05), pp. 870-873. Genova, Italy, Sept. 2005.

[3] Yin, P., Boyce, J., and Pandit, P. 2005. FMO and ROI Scalability. Joint Video Team (JVT) Document JVT-Q029. Nice, France, Oct. 2005.

[4] Stockhammer, T., Hannuksela, M.M., and Wenger, S. 2002. H.26L/JVT coding network abstraction layer and IP-based transport. Proceedings of International Conference on Image Processing, vol. 2, p. 485-488, 2002.

[5] Chen, Y., Xie, K., Zhang, F., Pandit, P., and Boyce, J. 2006. Frame loss error concealment for SVC. Journal of Zhejiang University SCIENCE A, 2006 7(5), p. 677-683.

[6] MPEG video group. 2003. Joint model of non-normative aspects of advanced video coding. ISO/IEC MPEG N5821, Aug. 2003.

[7] Kumar, S., Xu, L., Mandal, M.K., and Panchanathan, S. 2006. Error resiliency schemes in H.264/AVC standard. Elsevier J. of Visual Communication and Image Representation, vol. 17. April 2006.