

Adaptive Residual Interpolation: a Tool for Efficient Spatial Scalability in Digital Video Coding

Koen De Wolf, Davy De Schrijver, Wesley De Neve, and Rik Van de Walle
Multimedia Lab, Department of Electronics and Information Systems (ELIS),
Ghent University (UGent), Ghent, Belgium and IBBT, Belgium.

Abstract - Scalable video coding enables multimedia content providers to cope with the heterogeneity of networks and multimedia devices. In this paper, we introduce the novel concept of Adaptive Residual Interpolation (ARI) as a tool for efficient spatial scalable video coding. This new approach is based on the inter-layer prediction of residual data. We have implemented the ARI tool by making use of the Joint Scalable Video Model software. Tests have shown that our techniques for residual prediction result in bit rate savings of up to 8 % without diminishing visual quality. The ARI tool is especially useful for video sequences with low motion activity or when the residual data in the spatial enhancement layer is higher quantized than the data in the base layer.

Keywords: H.264/AVC, Scalable Video Coding, Inter-layer Prediction, Interpolation.

1 Introduction

The broad range of terminals and networks used in contemporary multimedia consumption chains poses some hard challenges for multimedia content distributors. In case of digital video, choices need to be made concerning which spatial resolutions, frame rates, and bit rates to support, hereby taking into account the properties of the targeted usage environment, i.e., the capabilities of the networks and devices used. Scalable Video Coding (SVC) can offer a solution for this challenge. To a certain extent, a scalable coded video bit stream can be automatically adapted so that it can be transported over the targeted networks and displayed on the targeted multimedia devices.

The traditional aim of video compression algorithms is to perform a reversible conversion of video data to a format that requires fewer bits. This conversion results in formatted data that can be stored or transmitted more efficiently. SVC adds an extra dimension to this goal, i.e., adaptivity. This enables multimedia content providers to cope with the heterogeneity of networks and multimedia devices. In general, scalable video schemes are able to encode a video sequence once, extract the suitable data for the targeted usage environment from the parent bit stream, and decode the extracted data. The three most common scalability axes are temporal, spatial, and SNR (Signal-to-Noise Ratio). A reduction of the quality along the temporal axis results in a decrease of the frame rate; along the spatial

axis in a smaller spatial resolution; and along the SNR axis in a lower visual quality. Ideally, every scalability axis has to be independently accessible from the encoded bitstream. Research is conducted on other types of scalability, e.g., complexity scalability, Region Of Interest scalability (ROI), color-depth scalability, etc.

In this paper, we introduce the novel concept of Adaptive Residual Interpolation (ARI) as a tool for efficient spatial scalability in digital video coding. This is a new approach in scalable video coding for the inter-layer prediction of residual data.

The remainder of this paper is organized as follows. First, we introduce the Joint Scalable Video Model (JSVM), as currently under development by the Joint Video Team (JVT). In the next section, we give a brief overview of the coding tools used in contemporary video compression schemes. Special interest goes out to the tools which can be used for enabling spatial scalable video coding. The subsequent section describes the concept of ARI, its benefits, complexity, and usage. In Sect. 4, a description of the conducted tests is given, as well as a discussion of the obtained results. We are winding up this paper with conclusions and remarks in Sect. 5.

2 Joint Scalable Video Model

The Moving Picture Experts Group (MPEG, ISO/IEC JTC1/SC 29/WG 11) and the Video Coding Experts Group (VCEG, ITU-T SG16) are currently exploring the field of SVC [1]. As the Joint Video Team, they started the development of new SVC algorithms in 2004. The resulting specification, commonly referred to as the Joint Scalable Video Model [2], is an extension of the non-scalable single-layer H.264/MPEG-4 Advanced Video Coding scheme (H.264/AVC) [3]. As a consequence, the Base Layer (BL) of the scalable bit stream should be H.264/AVC-compliant. In the next subsection, we elaborate on the overall structure of the JSVM. In Sect. 2.2, we discuss inter-layer redundancy and how the JSVM makes use of this redundancy.

2.1 JSVM Structure

The overall structure of a possible encoder, providing three spatial levels, is given in Fig. 1. In this figure, one can see that the original input video sequence has to be downscaled in order to obtain the different spatial layers (resulting in

spatial scalability). On each spatial layer, a temporal decomposition is performed. In the JSVM, this can be achieved by using hierarchical B-slice coded pictures [4] (see Sect. 2.2.1). This temporal decomposition leads to a motion vector field and residual texture data. The layered structure of the JSVM allows the use of motion and texture information of lower spatial layers for the prediction of this information in the higher layers. Finally, the texture information is spatially transformed and entropy coded by using Fine or Coarse Grain Scalability (FGS and CGS) to obtain SNR scalability.

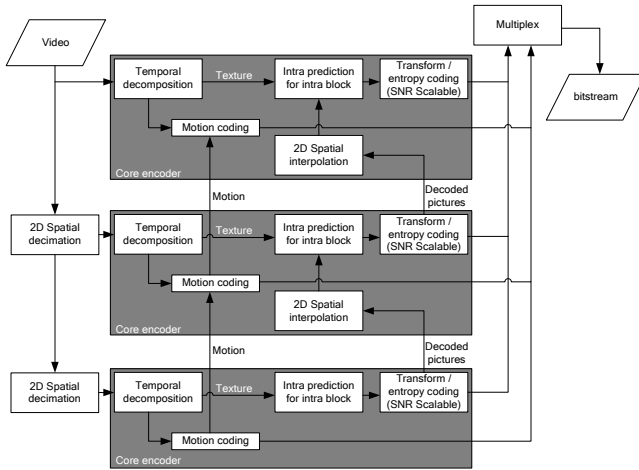


Fig. 1: JSVM encoder structure for providing three spatial levels [3]

The structure of a bit stream generated by the coding scheme as given in Fig. 1, is depicted in Fig. 2. Similar to H.264/AVC, a JSVM bit stream is a succession of Network Abstraction Layer Units (NALUs). Every NALU starts with a header which is succeeded by the actual payload. The NALU header contains the type of the unit. These NALUs can contain the coded data necessary for the reconstruction of luma and chroma samples or general information pertaining to the bit stream.

Every JSVM bit stream starts with a NAL Unit containing a Supplemental Enhancement Information (SEI) message with *payloadType* equal to 22. This metadata message contains information about the scalability axes incorporated in the bit stream such as the number of spatial levels, the temporal decomposition, the spatial resolution of the BL, the frame rate, inter-layer dependencies. Note that a decoder does not need this SEI message for the decoding of the bit stream itself.

After this SEI message, a number of NALUs, which embody Sequence Parameter Sets (SPSs), are contained in the bit stream. In particular, at least one SPS is needed for every spatial layer. An SPS is applicable to a complete sequence of pictures of a particular spatial layer and contains information about the profile used, the spatial resolution of the pictures in the sequence, etc. A number of NALUs containing Picture Parameter Sets (PPSs) are also encapsulated in the bit stream. A PPS applies to a number

of pictures of a sequence and contains information regarding the type of the entropy encoding that is being used, the number of slice groups, etc.

Finally, the NALUs containing slice information are integrated into the bit stream. This slice information is necessary to reconstruct the coded pictures. When the slice data of the unit belongs to an Enhancement Layer (EL), scalability information regarding the quality and temporal level will be present in the NALU header.

SEI messages with *payloadType* equal to 10 contain sub-sequence information of the next NALU. These messages will precede NALUs which are part of a sub-sequence and can be used to assist the extractor in generating temporally reduced bit streams from the H.264/AVC compliant layer (typically the base layer) [5]. For reductions along the spatial, temporal or quality axis, the extractor needs to parse the header of the NALUs which contain slice data. The mentioned SEI messages are very important to satisfy the requirement to support efficient bit stream extraction.

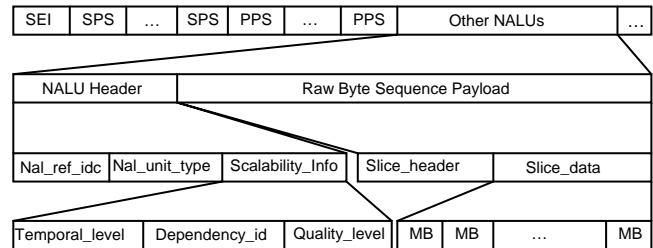


Fig. 2: Structure of a scalable bit stream

2.2 Inter-layer Redundancy

Video compression algorithms try to minimize the number of bits that are needed for the reproduction of the coded video at a predefined quality level. Reducing redundancy in the video source is the key concept for achieving this.

2.2.1 Temporal Redundancy

Motion Estimation and Motion Compensation (ME/MC) is considered as the main temporal decorrelation technique for inter-picture prediction in conventional video compression schemes (3D signals). Subsequent pictures in a video sequence are often very similar. We can use this property to remove inter-picture redundancy in order to achieve higher compression ratios. Efficient motion modeling assumes an accurate measurement of the displacement of pixels between two frames.

In contemporary video compression schemes, pictures are divided in blocks of pixels and use a block matching algorithm to estimate the displacement of a block between the current pictures and a previous coded picture. The displacement of this block will then be coded and stored in the resulting bit stream and is often referred to as the Motion Vector (MV) for that particular block. The best matching block will then be subtracted from the original one in the current picture. The difference between the latter

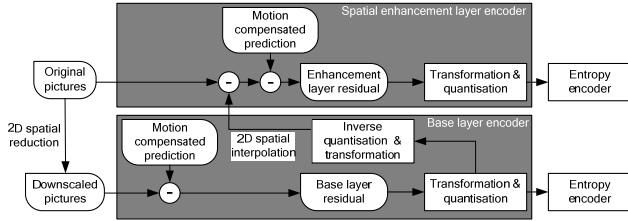


Fig. 4: Prediction of EL residual.

2.2.2.1 Intra-texture Prediction

When a MB of a picture in the BL is intra predicted, the corresponding block in the EL can also be predicted using techniques such as intra-macroblock prediction [2]. The intra-prediction mode of the block from the BL may serve as an estimation of the best intra prediction mode for the corresponding block in the EL.

2.2.2.2 Residual Prediction

When the motion vector of a particular block does not differ much from the motion vector of the corresponding block in a higher spatial layer, it is likely that after motion compensation the residuals for those blocks will show a high resemblance. We can use this feature to predict the residual for such blocks in higher resolution layers based on an upsampled version of the lower layer residual (see Fig. 4). Subtracting the interpolated residual of the BL from the original picture of the EL before ME/MC may result in lower prediction errors. In Fig. 5, the structure of the decoder using residual prediction is shown.

In the JSVM, a bilinear interpolation filter is defined for the upsampling of the residuals. When residual interpolation is used for the coding of a MB, `residual_prediction_flag` is set to 1 in the bit stream for that particular MB.

2.2.2.3 SNR

To allow quality scalability, the obtained residual pictures can be coded using a Fine-Granular Scalability (FGS) based coding mechanism. This results in a layered approach of progressive refined succeeding quality layers.

3 Adaptive Residual Interpolation

The contribution of this paper is the concept of Adaptive Residual Interpolation. We define ARI as a prediction method applied in the context of SVC which evaluates multiple interpolation filters in a rate-distortion optimized sense, used for the upsampling of BL residuals on a MB basis.

Our preliminary research has shown that the use of multiple predefined interpolation filters for the upsampling of base layer residuals is meaningful on a picture basis [13, 14]. However, higher gains in terms of bit rate and PSNR can be achieved when it is possible to switch between

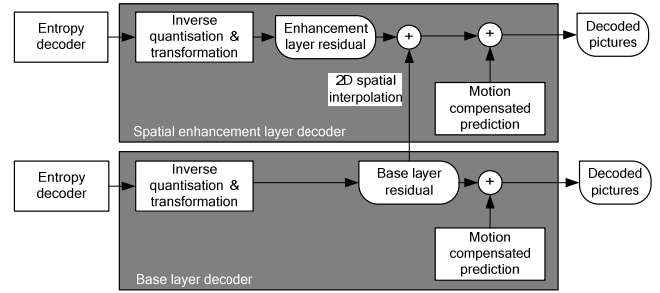


Fig. 5: Structure of the decoder using residual prediction

multiple predefined interpolation filters on MB level.

The amount of filters that will be evaluated for the interpolation of the BL residuals has a major impact on the time-complexity of the encoder. When all coding modes for a particular MB are evaluated, the maximum number of evaluations is given by

$$(\#Filters + 1) * (\#MB_{Inter_Modes} + \#MB_{Intra_Modes})$$

where $\#Filters$ denotes the number of filters used for residual interpolation, $\#MB_{Intra_Modes}$ denotes the number of modes that can be used for intra prediction and $\#MB_{Inter_Modes}$ denotes the number of modes that can be used for inter prediction. $+1$ denotes the coding mode when residual prediction is omitted.

Although selecting the best interpolation filter for every MB gives best coding results, it also constructs a larger search space. Moreover, the choice of which interpolation filter is used for which MB needs to be signaled in the bit stream. When we apply our technique to the JSVM, this can be done efficiently by using Context Adaptive Variable Length Coding (CAVLC) or Context Adaptive Binary Arithmetic Coding (CABAC).

3.1 Test setup

For the performance evaluation of the described prediction method, three sequences with different motion characteristics (ranging from low to high motion activity) were encoded. A short description with respect to the motion features of the tested sequences can be found in Table 1. All sequences consist of 300 pictures and contain two spatial layers which have QCIF (176x144) and CIF (352x288) resolution. The pictures of the low resolution sequence were generated using an 11-tap downsampling filter $(1, 0, -5, 0, 20, 32, 20, 0, -5, 0, 1)/32$ [15].

Only the first picture of a sequence was coded as an I-slice coded picture. P-slice coded pictures are inserted every 16 pictures. All pictures consist of exactly one slice. The coded bit stream has five dyadic temporal layers. Furthermore, four values for the Quantization Parameter (QP) of the BL and three values for the QP of the EL are used, resulting in nine test configurations per sequence. The tested combinations are shown in Table 2. The QPs listed in the table were chosen in order to support a broad range of possible use cases and applications.

In our tests, we have used two interpolation filters for the upsampling of the residuals: the bilinear filter (1, 1)/2 and the 6-tap interpolation filter used for sub-pixel motion estimation as defined in the H.264/AVC specification (1, -5, 20, 20, -5, 1)/32. These particular filters are already incorporated in the H.264/AVC SVC specification for other purposes. Therefore, no extra filters need to be designed.

The value of the newly introduced syntax element `adaptive_interpolation_filter` can embody 0 or 1. This syntax element is coded using the CABAC entropy coder for every MB which is coded using residual prediction (i.e., `residual_prediction_flag` = 1). Based on a statistical analysis of a test set, we have built up a context model for this syntax element. In our test setup, ARI is only enabled for P-slice coded pictures. The gain for ARI in B-slice coded pictures is marginal and not discussed here.

Table 1: Description of the tested sequences

Sequence name	Description
Foreman	Man presenting a construction yard; moderate movement from both the subject and the camera
mother & daughter	Mother and daughter sitting in front of a camera; no camera movement
Stefan	Tennis player; high motion from camera and subject, complex textures

Table 2 : Test configurations

Configuration	1	2	3	4	5	6	7	8	9
QP BL	0	12	12	12	24	24	24	36	36
QP EL	24	12	24	36	12	24	36	24	36

3.2 Discussion of the Results

In Fig. 6, the percentage of MBs that are coded using residual prediction is shown. fn on the X-axis denotes configuration n for the foreman sequence, mn configuration n for the mother & daughter sequence, and sn configuration n for the stefan sequence where n can be found in Table 2. The dots with drop lines represent the percentage of MBs in the EL whose residual is predicted with a single filter for the interpolation of the BL residual as defined in the current version of the JSVM [2]. The stacked columns give the amount of MBs when one out of two filters can be chosen for the residual interpolation. Note that these decisions, on whether to use residual prediction and on which filter to use, are made in a rate distortion optimized fashion. In particular, this means that all coding modes with full motion estimation are evaluated. From the figure, we see that more MBs use residual prediction when ARI is being used. This means that these MBs are coded more efficiently. We also see that on average about 45 % of the MBs which were coded using

residual prediction with one interpolation filter are now predicted with the other filter when ARI is being used. Again, this means that these MBs are coded more efficiently. When the QP of the BL is higher than or equal to the QP of the EL (configurations 2, 5, 6, 8, and 9), relatively few MBs are coded using residual prediction compared to the other configurations. This can be intuitively explained by the fact that most of the information contained in the BL residual is futile or lacking precision due to quantization and therefore it is more efficient to code the MBs discarding residual prediction. For example in configuration 8, all sequences are coded using ARI in less than 5 % of the MB.

In Figs. 7 to 15, we see that the use of ARI results in bit rate savings of up to 8 % per P-slice coded picture for the tested configurations without significant losses in terms of PSNR. Moreover, for about 38 % of the pictures the usage of ARI results in bit rate savings and PSNR gains.

If we look into more detail, we see that when the QP of the BL is lower than the QP of the spatial EL, almost all the pictures can be coded using fewer bits, while still maintaining visual quality. This is especially true when the QP of the BL is low. In that case more information is contained in the BL residual which can be used for inter-layer residual prediction. However, on Fig. 12, we see that even when quantization of the BL is high; its residual can still be used to optimize the coding of the spatial EL. In the latter case, bit rate savings are rather small.

One can also observe that ARI achieves higher bit rate savings for low motion sequences such as mother & daughter. This can be explained by the fact that the motion vectors from the spatial layers of this sequence are more aligned and hence the residuals of both spatial layers are more correlated. In sequences with complex motion, MBs in higher layers are usually further divided in sub-macroblocks which results in a more accurate motion vector field and a residual with lower energy. This can be coded more efficiently than using inter-layer residual prediction.

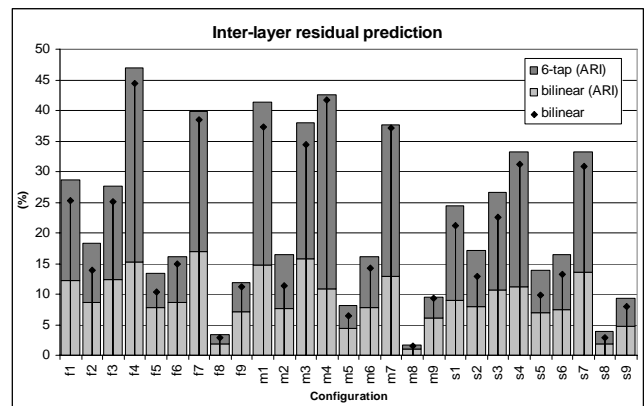


Fig. 6: Usage of inter-layer residual interpolation, expressed in %

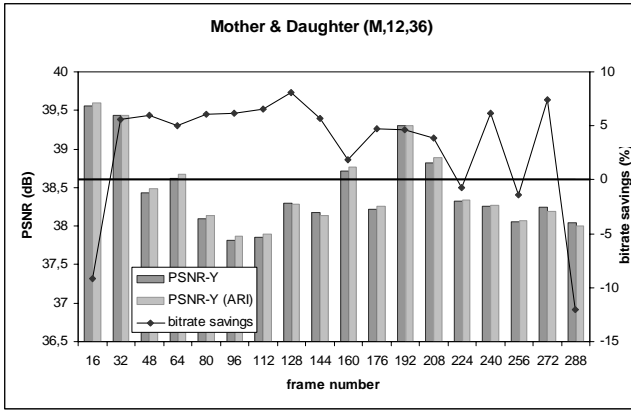


Fig. 7: Mother & daughter, QP BL = 12, QP EL = 36

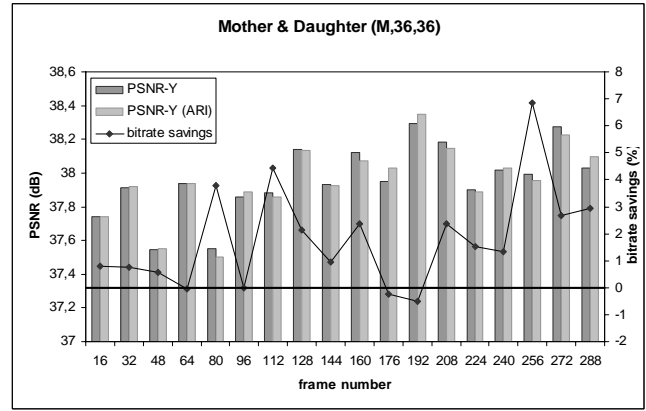


Fig. 9: Mother & daughter, QP BL = 36, QP EL = 36

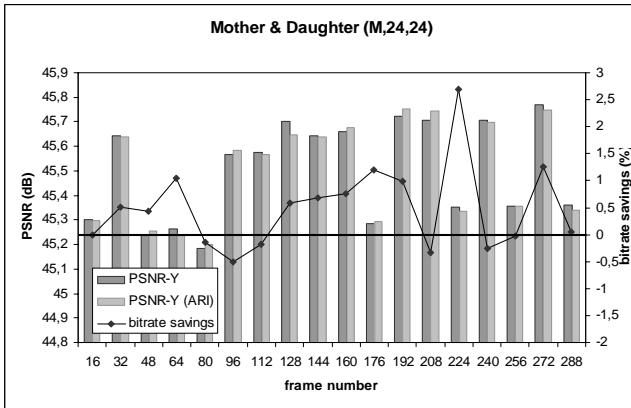


Fig. 8: Mother & daughter, QP BL = 24, QP EL = 24

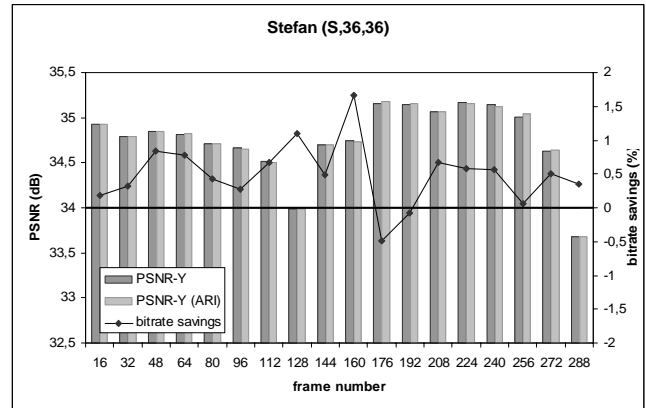


Fig. 10: Stefan, QP BL = 36, QP EL = 36

4 Conclusion and Remarks

In this paper, we discussed the feasibility of reusing residual texture information from the BL. We have described the novel concept of Adaptive Residual Interpolation (ARI) which can be used for the inter-layer prediction of residuals on a MB basis. From the conducted tests, we have seen that our residual prediction technique results in bit rate savings of up to 8 % without diminishing visual quality. For about 38 % of the P-slice coded pictures, the usage of ARI even results in bit rate savings and PSNR gains. Our tool is especially useful for video sequences with low motion activity or when the spatial EL is higher quantized than the BL.

5 Acknowledgement

The research activities that have been described in this paper were funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the fund for Scientific Research-Flanders (FWO-Flanders), the Belgian Federal Science Policy Office (BFSP), and the European Union.

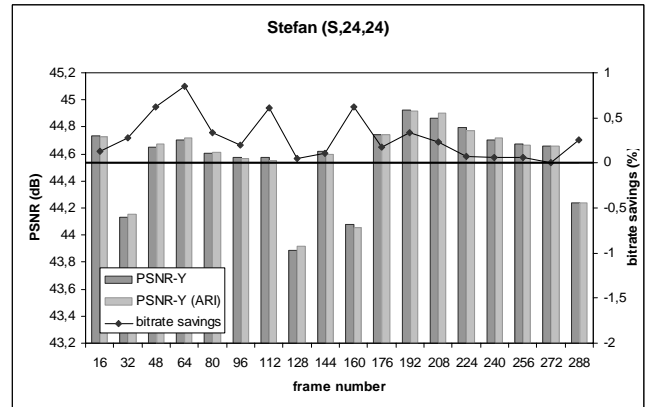


Fig. 11: Stefan, QP BL = 24, QP EL = 24

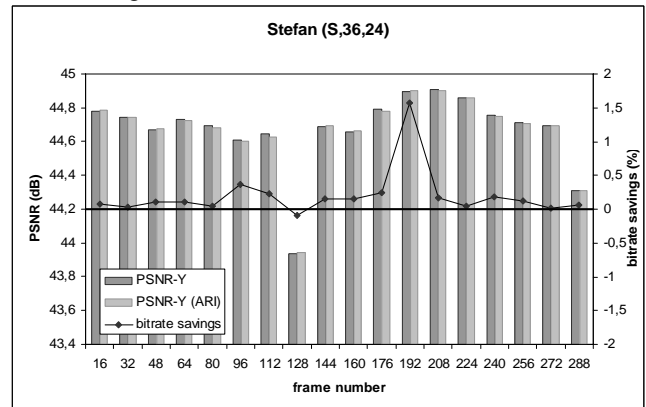


Fig. 12: Stefan, QP BL = 36, QP EL = 24

6 References

[1] ISO/IEC JTC1/SC29/WG11, "Applications and requirements for scalable video coding," ISO/IEC JTC1/SC29/WG11 N6880, Januari 2005.

[2] J. Reichel, M. Wien, and H. Schwarz, Eds., "Joint Scalable Video Model JSVM 4", Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, 2005.

[3] ITU-T and ISO/IEC JTC 1, "Advanced video coding for generic audiovisual services, ITU-T Rec. H.264 and ISO/IEC 14496-10 AVC (2003),".

[4] M. Hannuksela, "Enhanced concept of a GOP. doc. JVT-B042," January 2002.

[5] W. De Neve, D. Van Deursen, D. De Schrijver, K. De Wolf, R. Van de Walle, "Using bitstream structure descriptions for the exploitation of multi-layered temporal scalability in H.264/AVC's base specification," Lecture Notes in Computer Science 3767 (2005) 641–652.

[6] G. Sullivan and R. Baker, "Rate-distortion optimized motion compensation for video compression using fixed or variable size blocks," in Proceedings of the IEEE Global Telecommunications Conference, Phoenix, AZ, December 1991, vol. 3, pp. 85–90.

[7] M. Gothe and J. Vaisey, "Improving motion compensation using multiple temporal frames," in IEEE Pac. Rim Conference on Communications, Computers, and Signal Processing, 1993, vol. 1, pp. 157–160.

[8] T. Wiegand, X. Zhang, and B. Girod, "Block-based hybrid video coding using motion-compensated longterm memory prediction," in Proceedings of the Picture Coding Symposium, 1997.

[9] G. Van der Auwera, A. Munteanu, P. Schelkens, and J. Cornelis, "Bottom-up motion compensated prediction in the wavelet domain for spatially scalable video coding," in IEE Electronics Letter, 2002, vol. 38-21, pp.1251–1253.

[10] M. Mrak, G.C.K. Abhayaratne, and Ebroul Izquierdo, "Scalable generation and coding of motion vectors for highly scalable video coding," in Picture Coding Symposium 2004 (PCS-04), 2004.

[11] M. Mrak, G.C.K. Abhayaratne, and Ebroul Izquierdo, "On the influence of motion vector precision limiting in scalable video coding," in International Conference on Signal Processing (ICSP'04), 2004, vol. 2 of Proc. ICSP, pp. 1143–1146.

[12] S. A. Martucci, "Symmetric convolution and the discrete sine and cosine transforms," IEEE Trans. Sig. Processing, vol. 42, pp. 1038–1051, 1994.

[13] K. De Wolf, Y. Dhondt, J. De Cock, and R. Van de Walle, "Complexity analysis of interpolation filters for scalable video coding," in Proceedings of Euromedia, Toulouse, 4 2005, pp. 93–96.

[14] K. De Wolf, R. De Sutter, W. De Neve, and R. Van de Walle, "Comparison of prediction schemes with motion information reuse for low complexity spatial scalability," in Proceedings of SPIE/Visual Communications and Image Processing, Beijing, 7 2005, vol. 5960, pp. 1911–1920, SPIE.

[15] H. Schwarz, D. Marpe, and T. Wiegand, "Scalable extension of H.264/AVC," in ISO/IEC JTC 1/SC 29/WG11 MPEG2004/M10569/S03, 2004.

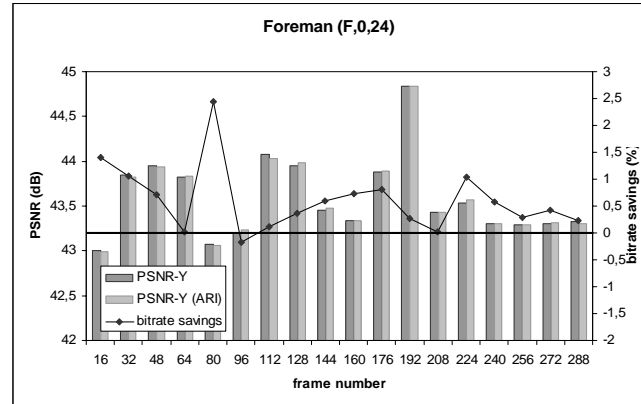


Fig. 13. Foreman, QP BL = 0, QP EL = 24

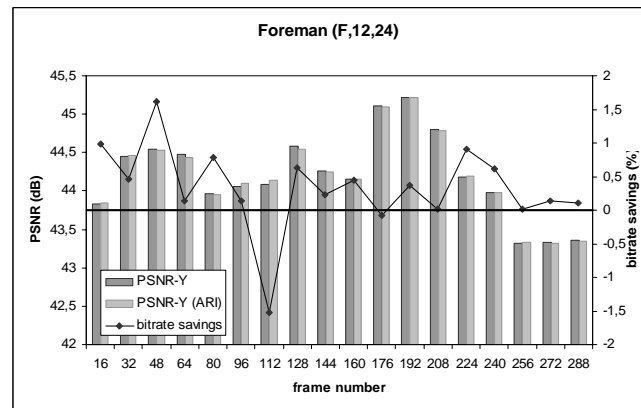


Fig. 14. Foreman, QP BL = 12, QP EL = 24

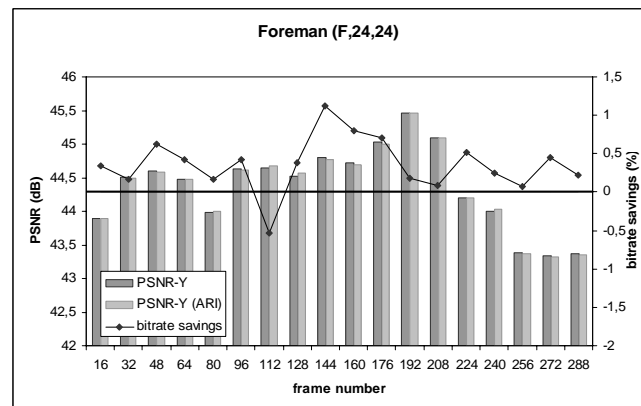


Fig. 15. Foreman, QP BL = 24, QP EL = 24