R-D OPTIMIZED MULTI-LAYER ENCODER CONTROL FOR SVC

Heiko Schwarz and Thomas Wiegand

Fraunhofer Institute for Telecommunications – Heinrich Hertz Institute,
Image Processing Department, Image Communication Group
Einsteinufer 37, 10587 Berlin, Germany, [hschwarz|wiegand]@hhi.fraunhofer.de

ABSTRACT

The scalable video coding (SVC) extension of H.264/AVC was recently standardized. The encoder control that is described in the Joint Scalable Video Model for SVC specifies a bottom-up process in which first the base layer and then the enhancement layer is encoded. The base layer is encoded conforming to H.264/AVC without considering its impact on enhancement layers, which limits the achievable enhancement layer coding efficiency. The losses relative to single-layer H.264/AVC coding are unevenly distributed between base and enhancement layers. In this paper, we present a multi-layer encoder control for SVC, by which we can trade off coding efficiency for base and enhancement layers and generally improve the efficiency of spatial and fidelity scalable coding.

Index Terms— video coding, standards

1. INTRODUCTION

The scalable video coding (SVC) amendment [1] of H.264/AVC [2] has recently been standardized. SVC allows partial transmission and decoding of a bit stream resulting in lower temporal or spatial resolutions or reduced fidelity. Temporal scalability can be efficiently provided using hierarchical prediction structures and did not require any changes to H.264/AVC. Spatial and fidelity scalability are supported via a layered coding approach.

In each layer, the concepts of motion-compensated prediction and intra prediction are employed as in single-layer H.264/AVC coding. Statistical dependencies between different layers are exploited by inter-layer prediction. For enhancement layers, an additional macroblock type is provided, which signals that prediction data are inferred from co-located blocks in the lower layer. If these co-located blocks are intra-coded, the prediction signal is the potentially up-sampled (for spatial scalable coding) reconstructed intra signal of the lower layer. Otherwise, the macroblock partitioning and the motion parameters are inferred from the lower layer blocks. For the H.264/AVC inter macroblock types, (scaled) lower layer motion vectors can also be used as replacement for the usual spatial motion vector predictor. With the usage of residual prediction, the enhancement layer macroblock before deblocking is constructed by adding the (up-sampled) lower residual and the enhancement layer residual to the inter prediction signal. For fidelity scalable coding, inter-layer prediction of the intra and the residual signal are performed in the transform coefficient domain.

As an important feature of SVC, the inter-layer prediction design allows decoding with a single motion compensation loop; a complete reconstruction of lower layer pictures is not required. Although the SVC design supports single-loop decoding, the encoder generally needs to be operated in multi-loop mode in order to avoid drift between encoder and decoder reconstruction for all layers of an SVC bit stream.

The encoder control for SVC that is recommended in the Joint Scalable Video Model (JSVM) [3] specifies a bottom-up process in which first the base layer and then the enhancement layer is encoded leading to an uneven distribution of the coding efficiency losses relative to single-layer H.264/AVC coding between base and enhancement layers. In this paper we present an r-d optimized multi-layer encoder control that allows to trade off base and enhancement layer coding efficiency and generally improves the efficiency of spatial and fidelity scalable coding. The underlying idea is related to the problem of bit allocation in dependent coding environments, which e.g. has been addressed in [4].

2. JSVM ENCODER CONTROL

The JSVM [3] encoder control specifies that encoder decisions for all layers be made in sequential order starting at the bottom layer. For each access unit, at first the coding parameters $\mathbf{p_0}$ for the base layer are determined following the widely-used Lagrangian approach [5],

$$\mathbf{p_0} = \arg\min_{\{\mathbf{p_0}\}} D_0(\mathbf{p_0}) + \lambda_0 \cdot R_0(\mathbf{p_0}), \qquad (1)$$

without considering their impact on the enhancement layers. $D_0(\mathbf{p_0})$ and $R_0(\mathbf{p_0})$ represent distortion and rate associated with selecting parameter vector $\mathbf{p_0}$. λ_0 is the Lagrange multiplier, which is determined based on the chosen quantization parameter QP_0. Similarly to the base layer, coding parameters $\mathbf{p_i}$ for each enhancement layer i are determined by

$$\min_{\{\mathbf{p_i} | \mathbf{p_{i-1}} \cdots \mathbf{p_0}\}} D_i(\mathbf{p_i} \mid \mathbf{p_{i-1}} \cdots \mathbf{p_0}) + \lambda_i \cdot R_i(\mathbf{p_i} \mid \mathbf{p_{i-1}} \cdots \mathbf{p_0}). \qquad (2)$$

given the already determined coding parameters $\mathbf{p_{i-1}}$ to $\mathbf{p_0}$ for the lower layers the enhancement layer i depends on.

While the base layer coding efficiency is basically identical to that of single-layer coding (minor losses may result from the mandatory usage of constraint intra prediction in SVC), there is usually a loss in coding efficiency for the enhancement layers. Furthermore, the effective reuse of the base layer data for enhancement layer coding is limited, because the chosen base layer coding parameters are optimized for the base layer only and are not necessarily suitable for efficient enhancement layer coding.

3. JOINT MULTI-LAYER CONTROL FOR SVC

In order to overcome the disadvantages of the JSVM encoding algorithm we developed an encoder control for spatial and fidelity scalable coding by joint optimization of base and enhancement layer coding parameter selection. The order in which the coding parameters are determined and the encoding process for the top layer are not modified relative to JSVM. However, for layers employed for inter-layer prediction of higher layers, the impact on the coding efficiency of dependent enhancement layers is taken into account. Without loss of generality, the modifications of the encoder control are described for a simple two-layer configuration; but they can be easily generalized for a multi-layer scenario.

In the two-layer scenario, all base layer decisions are based on the minimization of the modified cost functional

$$\min_{\{\mathbf{p_0},\mathbf{p}_1|\mathbf{p_0}\}} \begin{aligned}&(1-w)\cdot\left(D_0(\mathbf{p_0})+\lambda_0\cdot R_0(\mathbf{p_0})\right)+\\ &w\cdot\left(D_1(\mathbf{p}_1\,|\,\mathbf{p_0})+\lambda_1\cdot\left(R_0(\mathbf{p_0})+R_1(\mathbf{p}_1\,|\,\mathbf{p_0})\right)\right)\end{aligned}. \quad (3)$$

The first and second term of eq. (3) represent weighted costs for base and enhancement layer, respectively. The weighting factor $w \in [0; 1]$ controls the trade-off between base and enhancement layer coding efficiency. By setting w equal to 0, the encoder control becomes identical to the JSVM algorithm and the base layer coding efficiency is maximized. When w is equal to 1, the base layer parameters are only optimized for the enhancement layer coding without taking the reconstruction quality of the base layer into account.

With the general concept of eq. (3), the minimization proceeds over of the product space of $\mathbf{p_0}$ and $\mathbf{p_1}$. However, as will be shown in the remainder of this section, the optimization space can be significantly reduced. In the following, the actual modifications to mode decision, motion estimation, and the selection of transform coefficient levels for the base layer are described in more detail.

3.1. Mode decision

The macroblock modes, sub-macroblock modes, and intra prediction modes m_0 for the base layer are generally selected by minimizing

$$\min_{\{m_0\}} \begin{aligned}&(1-w)\cdot\left(D_0(m_0)+\lambda_0\cdot R_0(m_0)\right)+\\ &w\cdot\sum_k \min_{\{m_{1,k}\in M_k(m_o)\}} D_1(m_{1,k})+\lambda_1\cdot\left(R_0(m_0)+R_1(m_{1,k})\right)\end{aligned}. \quad (4)$$

To improve readability, the conditional expressions for the arguments of $D_1(\cdot)$ and $R_1(\cdot)$ have been neglected. The distortion terms $D_0(m)$ and $D_1(m)$ for the considered block are measured as the sum of squared differences (SSD) between original signal and reconstructed signal when coding the block with mode m. For spatial scalable coding, the base layer original is obtained by downsampling as specified in the JSVM. The rate terms represent the number of bits that are needed for encoding the considered block or macroblock with mode m including prediction modes, motion parameters, and transform coefficient levels. The parameter k is an index over the enhancement layer blocks, which depend on the considered block in the base layer.

It should be noted that the set of evaluated enhancement layer modes $M_{1,k}(m_0)$ can be restricted in a way that it only includes the mode or modes with the highest probability of employing inter-layer prediction. Note that the simulations results in sec. 4 have been generated with the following restrictions. For fidelity scalability, the set $M_{1,k}(m_0)$ consisted only of the macroblock type, for which all prediction parameters are inferred. For spatial scalability, the mode with the same partitioning as the one inferred from the base layer but different motion parameters and the mode or modes with the next finer partitioning are additionally included.

3.2. Motion estimation

The motion vectors $\mathbf{v}_0$ for inter-coded base layer blocks are determined by minimizing the Lagrangian cost functional

$$\min_{\{\mathbf{v}_0\in S\}} \begin{aligned}&(1-w)\cdot D_0(\mathbf{v}_0)+w\cdot D_1(s\cdot\mathbf{v}_0)+\lambda^*\cdot R_0(\mathbf{v}_0)\\ &\text{with}\qquad \lambda^*=(1-w)\cdot\lambda_0+w\cdot\lambda_1\end{aligned}. \quad (5)$$

The distortions $D_0(\mathbf{v})$ and $D_1(\mathbf{v})$ are measured as sum of absolute differences (SAD) between the original signal and the prediction signal that is obtained by employing the motion vector $\mathbf{v}$. $R_0(\mathbf{v})$ specifies the number of bits needed for transmitting the motion vector $\mathbf{v}$. The parameter s is a scaling factor for the motion vector. It is equal to 1 for fidelity scalability and equal to 2 for dyadic spatial scalability.

In order to keep the motion estimation process simple, we only consider enhancement layer modes that reuse the (scaled) base layer motion vector; potential motion refinements in the enhancement layer are neglected at this point. For SNR scalable coding, the motion search can be further simplified by weighting the reference frames for base and enhancement layer instead of the distortion measures.

3.3. Selection of transform coefficient levels

Transform coefficient levels l are typically determined via forward transformation of the residual signal of a block and a following scalar quantization according to

$$l = \mathrm{sign}(t)\cdot\lfloor(\mathrm{abs}(t)+f\cdot q)/q\rfloor \quad (6)$$

with t representing a transform coefficient and q being the quantization step size. f is a control parameter for considering the non-uniform distribution of transform coefficients. The chosen value for f usually is inside the interval [0; 0.5].

Since we want to select the base layer transform coefficient levels l_0 in way that both the coding efficiency of base

and enhancement layer are taken into account, it is not possible to apply a simple quantization rule similar to (6). Instead, we employ a trellis-based approach, which is related to the scheme presented in [6]. But instead of optimizing transform coefficient levels for single-layer coding, we use the trellis for considering the impact on inter-layer residual prediction for the enhancement layer and employ a simplified sub-optimal search strategy. As illustrated in Fig. 1, the stages of the trellis are given by the scanning positions of the considered transform block, and the nodes of the trellis represent potential values for the transform coefficient levels.

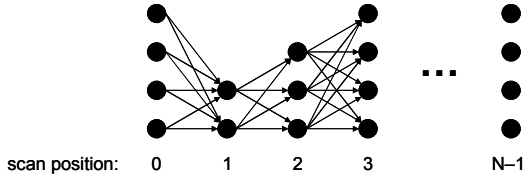

Fig. 1. Trellis representation for the transform coefficients levels of a block.

The problem of determining base layer transform coefficient levels for a transform block while considering their impact on enhancement layer coding efficiency has been converted into the problem of finding the path in the trellis, which is associated with the smallest rate-distortion cost. The cost measure for a path is given by the functional

$$(1-w)\cdot\left(D_0(\mathbf{l}_0)+\lambda_0\cdot R_0(\mathbf{l}_0)\right)+ w\cdot\sum_k D_1(\mathbf{l}_{1,k}\,|\,\mathbf{l}_0)+\lambda_1\cdot\left(R_0(\mathbf{l}_0)+R_1(\mathbf{l}_{1,k}\,|\,\mathbf{l}_0)\right) \quad (7)$$

with $\mathbf{l}_0$ and $\mathbf{l}_{1,k}$ representing the vector of scanned transform coefficient levels for the base and enhancement layer block, respectively. The distortions $D_0(\mathbf{l})$ and $D_1(\mathbf{l})$ are measured as SSD between the original and the reconstructed residual. The rates $R_0(\mathbf{l})$ and $R_1(\mathbf{l})$ represents the number of bits that are required for coding the block of transform coefficient levels $\mathbf{l}$. For each possible vector of base layer levels $\mathbf{l}_0$, the corresponding enhancement layer levels $\mathbf{l}_{1,k}$ are determined by forward transformation of the enhancement layer residual that is obtained after inter-layer residual prediction and a subsequent quantization according to eq. (6). Additionally, transform coefficient levels for the case without inter-layer residual prediction are determined via a forward transformation and a subsequent quantization according to (6) for both, base and enhancement layer. And finally, for each macroblock, the set of transform coefficient levels (optimized for the case with or without residual prediction) that results in the minimum cost measure (7) is selected.

In order to limit the complexity of the described approach, we introduced the following two simplifications:

1. We first determine the base layer transform coefficient levels $\mathbf{l}_{\min,0}$ and $\mathbf{l}_{\min,1}$ that minimize the distortion of base and enhancement layer residual, respectively. These levels are obtained via forward transformation and quantization according to (6) with $f = 0.5$. For spatial scalable coding, an additional downsampling operation is applied before the forward transformation in order to determine the levels $\mathbf{l}_{\min,1}$. The minimum and maximum values of transform coefficient levels for each scanning position can then be set equal to $\min(0, l_{\min,0}, l_{\min,1})$ and $\max(0, l_{\min,0}, l_{\min,1})$, respectively.
2. We employ a sub-optimal search strategy. At the beginning, all levels are set equal to 0, and then the trellis is evaluated stage by stage similar to the Viterbi algorithm. After determining the r-d costs for a sub-path from stage 0 to stage k (assuming transform coefficient levels equal to 0 for stages $n > k$), we only keep three sub-paths for further evaluation: The minimum cost path with a level equal to 0 at stage k, the minimum cost path with an absolute value equal to 1 for the level at stage k, and the minimum cost path with an absolute value greater than 1 for the level at stage k.

4. SIMULATION RESULTS

The impact of choosing the weighting factor w is demonstrated in Fig. 2 for the example of fidelity scalable coding with a simple IPPP coding structure. With the JSVM encoder control corresponding to $w = 0$, the base layer coding efficiency is virtually identical to that of single-layer coding, but a 1.5 dB PSNR loss relative to single-layer coding can be observed for the enhancement layer. By increasing the weighting factor w, the coding efficiency for the enhancement layer can be improved while reducing base layer coding efficiency. For w=0.75, the PSNR loss at the enhancement layer is reduced to 0.6 dB while incurring a base layer PSNR loss of 0.5 dB. For the case $w = 1$ the enhancement layer coding efficiency is similar to that of single layer coding, but the quality of the base layer is unacceptable, since its reconstruction quality is not controlled during encoding.

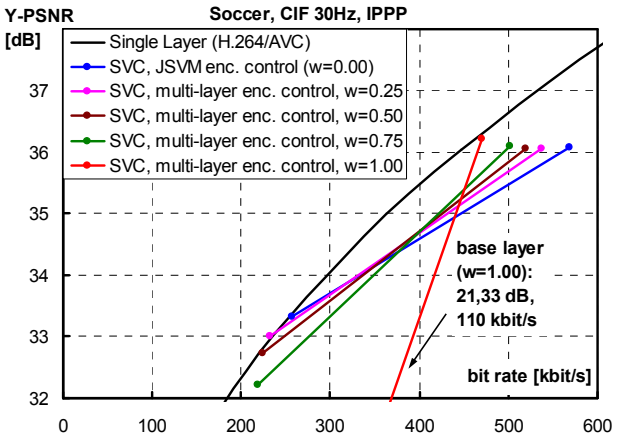

Fig. 2. Impact of choosing the weighting factor w on coding efficiency for fidelity scalable coding.

In a further experiment we used the developed multi-layer encoder control for optimizing fidelity and spatial scalable coding with SVC in a way that the coding efficiency

losses against single-layer H.264/AVC coding are nearly evenly distributed between base and enhancement layer. Typical simulation results for fidelity and dyadic spatial scalable coding are presented in Figs. 3 and 4. In both we chose a coding structure with hierarchical B pictures and a group of 16 pictures, which does not only provide dyadic temporal scalability with 5 temporal levels, but also increases coding efficiency for both single-layer and scalable coding. For fidelity scalable coding, the weighting factor was set equal to 0.25 for the pictures of the coarsest temporal resolution and equal to 0.75 for the remaining pictures. For spatial scalable coding, a weighting factor of 0.5 was used for all pictures.

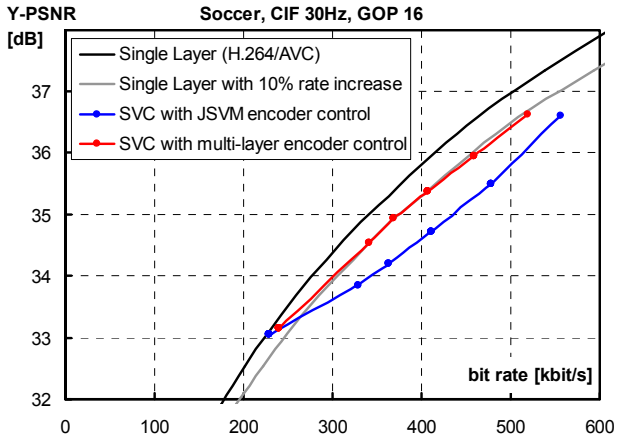

Fig. 3. Comparison of the JSVM and the proposed encoder control for fidelity scalable coding.

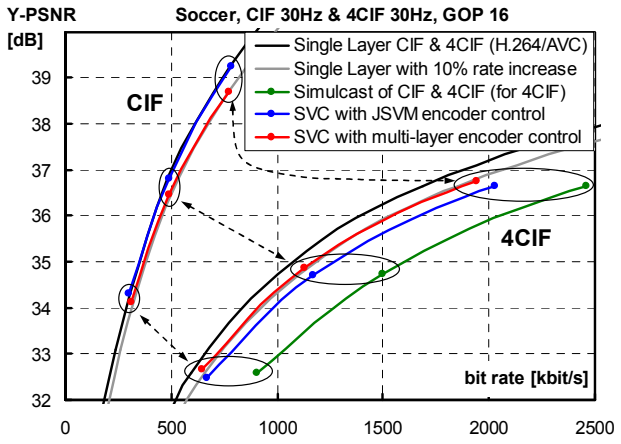

Fig. 4. Comparison of the JSVM and the proposed encoder control for spatial scalable coding.

All points of the rate-distortion curves for fidelity scalable coding in Fig. 3 were extracted from a single SVC bit stream by successively discarding fidelity enhancement representations starting with the finest temporal refinement level. By replacing the JSVM encoder control with the developed algorithm, the base layer quality became slightly worse, but a significant increase in coding efficiency could be observed for all enhancement layer points corresponding to 0.6 dB PSNR improvement at 400 kbit/s.

Similar results were also obtained for spatial scalable coding. For a better evaluation of spatial scalable coding, the diagram in Fig. 4 additionally shows the coding efficiency of simulcast (corresponding to a 40 % bit-rate overhead on average). Dashed arrows connect base and enhancement layer points that are included in the same bit-stream. Moreover, it should be noted that in addition to trading off base and enhancement layer coding efficiency also an increase of the efficiency of scalable coding is seen, which can be expressed by the percentage of the base layer rate that is reused for enhancement layer coding. The average base layer usage for spatial scalable coding has been increased from 64% for JSVM to 81% for the multi-layer encoder control.

Another important thing to note is that the simulation results show that the SVC can provide a suitable degree of scalability at the cost of a bit rate increase of approximately 10% relative to single-layer H.264/AVC coding for all representations included in a scalable bit stream.

5. CONCLUSION

We presented an r-d optimized multi-layer encoder control for SVC, which makes it possible to trade off base layer and enhancement layer coding efficiency and to generally improve the effectiveness of fidelity and spatial scalable coding. It could further be shown that SVC is capable of providing a reasonable degree of fidelity and spatial scalability with a small bit rate increase of about 10% relative to single-layer H.264/AVC coding. Future research is required to reduce the complexity of the presented encoding algorithms.

6. REFERENCES

[1] T. Wiegand, G. J. Sullivan, J. Reichel, H. Schwarz, and M. Wien, "Joint Draft 10 of SVC Amendment," *Joint Video Team*, doc. JVT-W201, San Jose, CA, USA, Apr. 2007.

[2] ITU-T Rec. | ISO/IEC 14496-10, "Advanced video coding for generic audiovisual services," version 3, 2005.

[3] J. Reichel, H. Schwarz, and M. Wien "Joint Scalable Video Model 10 (JSVM 10)," *Joint Video Team*, doc. JVT-W202, San Jose, CA, USA, Apr. 2007.

[4] K. Ramchandran, A. Ortega, and M. Vetterli, "Bit allocation for dependent quantization with application to multiresolution and MPEG video coders," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 533-545, Sep. 1994.

[5] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 688-703, July 2003.

[6] J. Wen, M. Luttrell, and J. Villasenor, "Trellis-based r-d optimal quantization in H.263," *IEEE Transactions on Image Processing*, vol. 9, no. 8, pp. 1431-1434, Aug. 2000.