Quality Scalable Low Delay Video Coding using Leaky Base Layer Prediction

Steffen Kamp and Mathias Wien
Institute of Communications Engineering
RWTH Aachen University, 52056 Aachen, Germany
E-mail: kamp@ient.rwth-aachen.de

***Abstract*—In this paper quality scalable video coding using leaky base layer prediction for low-delay applications is considered. The temporal prediction reference for the base layer is generated by calculating a weighted average of the quality base layer and quality enhancement layer reference pictures. This provides close to single layer performance at the enhancement layer rate point while introducing drift into the base layer if the enhancement layer is truncated or lost. Approaches using global weighting and locally adaptive weighting are investigated. The paper proposes an alternative approach for SNR scalable video coding targeted at applications usually close to the enhancement layer rate as operating point while providing good error recovery if enhancement layer packets are lost.**

I. Introduction

In SVC (scalable video coding) [1], the scalable extension to H.264|AVC [2] currently under development by the Joint Video Team (JVT) of ISO/MPEG and ITU, quality scalability is achieved using either coarse grain scalable (CGS) or fine grain scalable (FGS) enhancement layers (EL) on top of a independently coded closed-loop, AVC compatible base layer (BL) bitstream. For this paper we focus on IPPP coding, prohibiting prediction from future frames and therefore limiting the end-to-end delay. SVC coding of P slices with FGS quality scalability is done by adding progressive refinement (PR) slices to the base layer signal. PR for P-slices is coded using one of the two following methods: (a) The refinement signal is coded without temporal prediction on top of the base layer reconstructed picture, similar to MPEG-4 FGS [3] (see Figure 1). (b) A temporal prediction for the enhancement layer is used, where the prediction signal is computed using an adaptively weighted average of the base and enhancement layer reference pictures (adaptive reference, AR) [4] (see Figure 2), introducing a second motion compensation loop for the enhancement layer.[1] While PR slices with AR significantly improve the coding efficiency at the enhancement layer rate point, both FGS schemes favour the base layer rate point which is at single-layer performance, while the enhancement layer performance is well below single-layer performance.

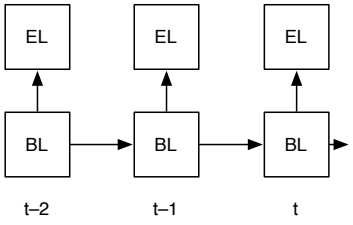

Fig. 1. MPEG-4 like FGS coding: The enhancement layer does not use any temporal prediction.

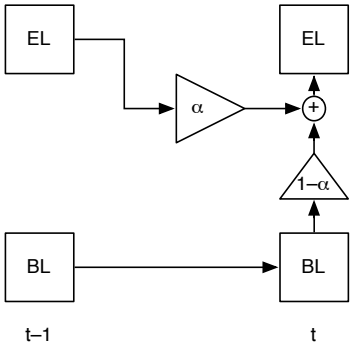

Fig. 2. PR coding using adaptive references (AR): The enhancement layer reference is computed as a weighted average of the base layer and enhancement layer reference pictures.

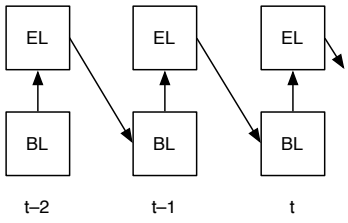

Fig. 3. Base layer prediction from the enhancement layer signal as in MPEG-2 SNR-Scalable profile.

The main reason for this behaviour is the sub-optimum temporal prediction in the base layer which only uses base layer reconstructed reference pictures. MPEG-2 SNR scalability [5] uses the full enhancement layer signal for temporally predicting the base layer (see Figure 3). While this yields almost single-layer performance if the enhancement layer signal is available, decoding the base layer in the absence of an EL signal is subject to temporal drift. In this paper, we propose to employ leaky base layer prediction, i. e. using a

[1]This describes only the principal approach. Refer to [4] for a detailed description.

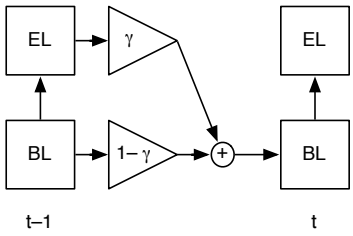

Fig. 4. Proposed approach: Generating the base layer reference as a weighted average of the base layer and enhancement layer reference pictures.

weighted average of reconstructed enhancement and base layer pictures for the temporal prediction of the base layer. This way the prediction quality can be significantly improved, yielding performance gains if the enhancement layer signal is available to the decoder while confining drift if the enhancement layer is truncated or lost. Similar approaches have been presented in [6] where the reference pictures are generated by decoding the enhancement layers up to a specific bit-plane and [7] where the partial enhancement layer signal is attenuated by a leak factor for computing the base layer prediction reference. In contrast to [7] and [8] where separate coding loops are used for each layer we only use a single motion compensation loop for base layer coding while enhancement layers are coded without further inter prediction.

This paper is organised as follows. Section II describes the leaky base layer prediction scheme. Complexity considerations are provided in Section III, while Simulation results and a discussion are presented in Section IV. Section V concludes this paper.

II. Description

The proposed leaky base layer prediction scheme generates the temporal prediction reference for motion compensated base layer coding by accessing both, reconstructed base and enhancement layer representations of the reference picture. A weighted average of both representations is calculated in the spatial domain. Let $\mathbf{r}_b^{t-1}$ be the base layer reference signal and $\mathbf{r}_e^{t-1}$ the enhancement layer reference signal for coding the current base layer block and taking necessary motion compensation into account. Then the final base layer reference signal for the coded picture at time index t, $\mathbf{r}_a^t$ is calculated as (see Figure 4)

$$\mathbf{r}_a^t = \gamma \cdot \mathbf{r}_e^{t-1} + (1-\gamma) \cdot \mathbf{r}_b^{t-1}, \quad 0 \leq \gamma \leq 1. \tag{1}$$

The inclusion of enhancement layer information for the base layer prediction obviously leads to temporal drift at the decoder if parts of the enhancement layer bitstream are lost during transmission. Therefore, this scheme is mainly suited for applications where it can be expected that the enhancement layer is retained most of the time and enhancement layer packet truncation or loss is the exception.

The leak factor γ provides for a trade off between enhancement and base layer prediction, allowing to reduce the decoder side drift at the cost of reduced coding efficiency.

The presented scheme provides a flexible method for quality scalability which is currently not considered in SVC. It conceptually encompasses MPEG-2 SNR scalability (Figure 3) when choosing $\gamma = 1$ for all inter macroblocks as well as MPEG-4 FGS coding for $\gamma = 0$ (Figure 1). Additionally it allows for a flexible selection of γ.

As an alternative to using a fixed value of γ in the reference signal generation for all macroblocks, we have also examined a local adaptation of γ. The adaptation is dependent on the macroblock coding mode of the currently coded base layer macroblock similar to the local adaptation of the leak factor for AR coding discussed in [9]. Each macroblock coding mode is assigned a fixed value of γ that is known to both encoder and decoder. Therefore, when using locally adaptive values of γ the mode decision should apply the respectively assigned value of γ for the evaluation of different coding modes.

In this paper we specifically used a set up with $\gamma = 0$ for all macroblocks not coded in Skip mode, while using a fixed γ for Skip blocks. The motivation for this approach is that the Skip mode is typically chosen when the reference signal is highly correlated with the current block. We therefore assume that Skip blocks are less likely to be the origin of temporal drift compared to blocks with higher residual signal energy.

It is important to note that while the proposed scheme modifies the base layer coding by feeding in enhancement layer information, no changes are required to the actual AVC base layer syntax or semantics. A possibility of integrating the enhancement layer prediction into SVC would be to signal the usage of this method through a flag in the suffix NAL unit [1] which is a SVC-only NAL unit carrying scalability information for base layer NAL units. Accordingly, the base layer bitstream remains decodable by existing standard AVC compliant decoders, which is an important design feature of SVC. However, being unaware of the enhancement layer signal, the base layer decoder suffers from temporal drift. If the enhancement layer signal is not lost during transmission, an appropriate SVC decoder using the proposed prediction scheme is capable of decoding a drift free signal.

The investigated approach can be combined with PR coding using AR [4], e.g. performing temporal prediction for all enhancement layer blocks or alternatively only for those blocks where $\gamma = 0$. Another possibility is to use the presented scheme for the first quality layer while using PR coding with or without AR for all higher quality layers. This configuration allows to use the bitrate at the first enhancement layer as operating point where almost single-layer performance is achieved while providing for dynamic adaptation to increased or decreased channel bandwidths. However, these ideas are not investigated in this paper and are subject to further study.

In the latest SVC draft, PR slices have been removed from the specification. However, our proposed scheme can still be implemented for CGS enhancement layers in SVC. Employing our proposed leaky prediction scheme would allow for a comparable trade off as presented for FGS coding in this paper.

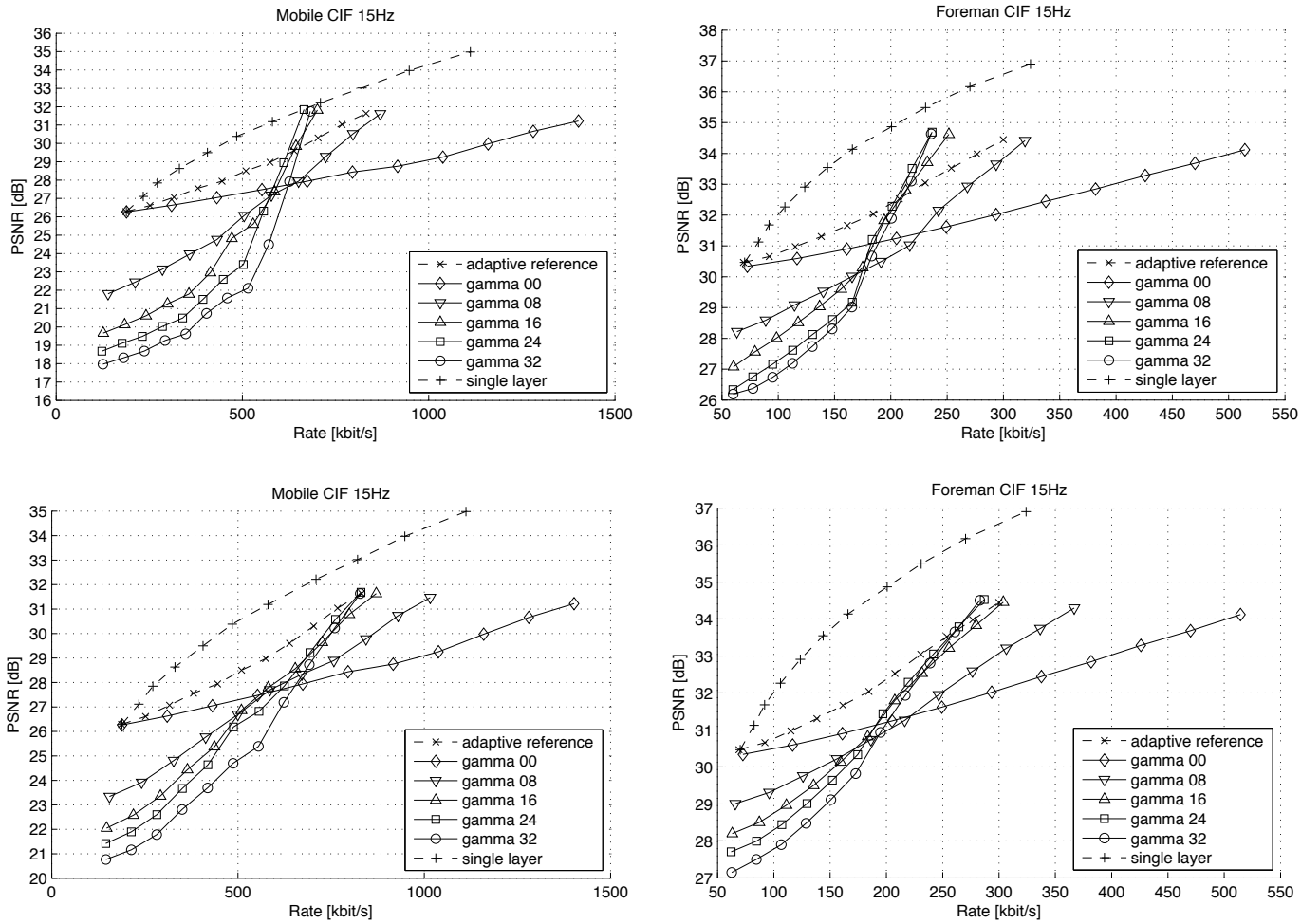

Fig. 5. Rate-distortion results at base layer QP $= 38$. Top: all macroblocks using leaky base layer prediction. Bottom: Leaky base layer prediction for skip blocks only and MPEG-4 like FGS for other macroblock coding modes.

III. Complexity Consideration

From a complexity perspective, the investigated method requires the same number of frame buffers as PR slices using AR. The prediction reference signals can be computed before performing motion compensation and only the base layer reconstruction employs a motion compensated coding loop with the weighted prediction signal $\mathbf{r}_a^t$. The prediction signal for coding the enhancement layer is then taken from the corresponding base layer picture $\mathbf{r}_b^t$. For PR slices with AR, the base layer uses motion compensated prediction from the base layer reference signal $\mathbf{r}_b^{t-1}$. Additionally, the enhancement layer uses motion compensated prediction from the enhancement layer reference signal $\mathbf{r}_e^{t-1}$. So while the computation of the weighted enhancement reference signal can also be performed before motion compensation, the enhancement layer still requires another motion compensated coding loop. Therefore, for our proposed scheme and for coding one enhancement layer, the motion compensation/interpolation complexity is roughly halved when compared to PR slices with AR (not taking memory accesses into account).

Generally, when using locally adaptive values of γ depending on the macroblock coding mode and precomputing the weighted reference pictures, one frame buffer for each unique value of γ is required. Alternatively, the weighted addition could be performed after motion compensation of base and enhancement layer reference pictures for each block, using two motion compensated prediction loops and therefore nullifying the complexity advantage of the proposed scheme. However, with the specific configuration of using only two unique values of γ as examined in this paper, the prediction references can be stored in the available base and enhancement layer frame buffers. Motion compensation for each block can then selectively use either of the two references, retaining the aforementioned complexity reduction.

IV. Results

Simulation results are provided for Foreman and Mobile sequences at CIF resolution and 15 Hz. All pictures except the initial I-picture are coded as P-pictures using the proposed leaky prediction scheme. Obviously, when using this coding structure temporal drift propagates through all frames,

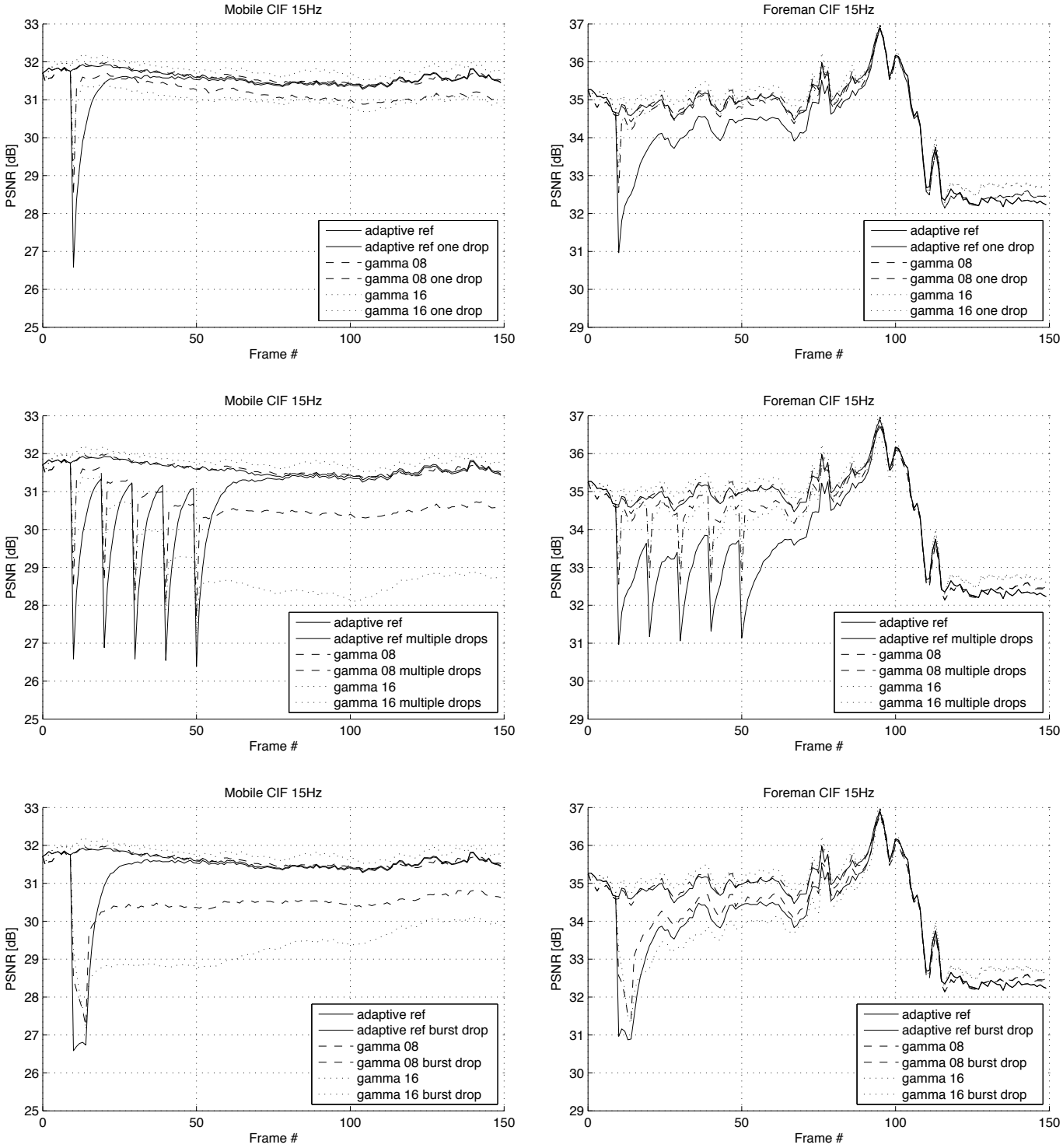

Fig. 6. Error recovery at $\gamma = \frac{1}{4}$ and $\gamma = \frac{1}{2}$ compared to AR coding when dropping enhancement layer packets at frame 10 (one drop, top), frames 10, 20, 30, 40, 50 (multiple drops, middle) and frames 10, 11, 12, 13, 14 (burst drop, bottom). For all cases, the base layer was received without loss.

especially for MOBILE which is coded using relatively long prediction chains due to its regular motion characteristics. For such cases, further limiting of the temporal drift can be achieved by coding additional I pictures. In order to prevent spatial drift in intra coded blocks of P pictures, we used constrained intra prediction for our simulations, which restricts intra prediction reference samples to previously intra coded samples.

For comparison we provide simulation results for single-layer (non-scalable) coding and PR coding using AR (labelled *adaptive reference*) using the JVT core experiment testing conditions for QP $= 38$ and one FGS layer in [10]. Results for $\gamma = 0$ are identical to plain FGS coding without temporal prediction in the EL.

In our current implementation, the coded value of γ is quantised to 32 possible levels. The legends of rate-distortion plots show the coded value $\gamma_{\mathrm{coded}} = 32 \cdot \gamma$. Rate points were achieved by successively truncating enhancement layer packets evenly across all coded pictures in 10 % decrements.

When using leaky base layer prediction for all macroblocks the performance at the enhancement layer rate is almost identical to single-layer coding for MOBILE, while for FOREMAN PSNR is about 1 dB below single-layer (see Figure 5). Results for using enhancement layer references only for Skip blocks expose less drift at the base layer rate compared to using prediction from the enhancement layer for all inter coded blocks. The performance at the enhancement layer rate is typically reduced. For some scenarios this results in improved overall rate-distortion behaviour: FOREMAN with $\gamma = \frac{1}{2}$ for skip-blocks only (*gamma 16* in Figure 5 bottom right) outperforms $\gamma = \frac{1}{4}$ for all inter coded blocks (*gamma 08* in Figure 5 top right) over the whole range of bitrates. For γ around 0.5 both methods provide enhancement layer performance comparable to PR coding with AR while only requiring a single decoding loop.

Due to the potentially large amount of drift at the base layer point, the usage of the presented method should be restricted to near constant-quality scenarios. In this case, it is expected that the decoder mostly receives the complete or almost complete enhancement layer information but some enhancement packets may be lost. For such a scenario the quality at the target rate (highest rate point) will be improved over AR-PR coding and close to single-layer while the base layer would only serve as low-quality but still decodable fall back.

To evaluate packet loss cases we have examined three packet loss scenarios (see Figure 6): (a) Only the enhancement layer packet of frame 10 is lost (*one drop*). (b) Enhancement layer packets of frames 10, 20, 30, 40 and 50 are lost (*multiple drops*). (c) Enhancement layer packets of frames 10, 11, 12, 13 and 14 are lost (*burst drop*). Error recovery from a single dropped EL packet is almost instant for our leaky prediction scheme while requiring a significant amount of frames in case of AR coding. While for FOREMAN the PSNR approaches the loss-less performance after the packet drops, PSNR does not fully recover for MOBILE due to longer prediction chains, especially for multiple drops. Subjectively, we found the degradation of visual quality not as significant as the PSNR drops in Figure 6 suggest.

V. CONCLUSION

We have shown a method for achieving quality scalability in SVC that is not currently available in the SVC draft. While allowing temporal drift in the base layer, the coding efficiency for higher rate points can be improved. Results for two specific implementations of enhancement layer prediction for the base layer have been provided. In contrast to MPEG-4 FGS or PR coding with AR in SVC the presented approach benefits the higher rate points while penalising lower qualities. An advantage of the scheme is that error recovery after FGS packet loss is almost instant. Although further investigations are required, this method could be considered as a valuable approach for scalable video coding schemes, giving the encoder an additional degree of flexibility.

REFERENCES

[1] Thomas Wiegand, Gary Sullivan, Julien Reichel, Heiko Schwarz, and Mathias Wien, "Joint draft 7 of SVC amendment," Doc. JVT-T201, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, 20th Meeting, Klagenfurt, Austria, July 2006.

[2] *ITU-T Rec. H.264 & ISO/IEC 14496-10 AVC: Advanced Video Coding for Generic Audiovisual Services*, version 3: 2005.

[3] ISO/IEC 14496-2, "Information technology – Generic coding of audio-visual objects: Visual," 1998.

[4] Yiliang Bao and Marta Karczewicz, "CE7 report, FGS coding for low-delay applications," Doc. JVT-Q039, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, 17th Meeting, Nice, France, Oct. 2005.

[5] ISO/IEC 13818-2, "Information technology – Generic coding of moving pictures and associated audio information: Video," 1996.

[6] Mihaela van der Schaar and Hayder Radha, "Motion-compensation based fine-granular-scalability (MC-FGS)," Doc. MPEG2000/M6475, ISO/IEC JTC1/SC29/WG11, La Baule, France, Oct. 2000.

[7] Hsiang-Chun Huang, Chung-Neng Wang, and Tihao Chiang, "A robust fine granularity scalability using trellis-based predictive leak," *IEEE Transactions on Circuits and Systems for Video Technology*, June 2002.

[8] Sangeun Han and Bernd Girod, "Robust and efficient scalable video coding with leaky prediction," in *Proc. IEEE Int. Conference on Image Processing ICIP '2002*, 2002, vol. 2, pp. 41–44.

[9] Steffen Kamp and Mathias Wien, "Local adaptation of leak factor in AR-FGS," Doc. JVT-S092, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, 19th Meeting, Geneva, Switzerland, Apr. 2006.

[10] Seyoon Jeong, Steffen Kamp, Xianglin Wang, and Xiangyang Ji, "CE5: Improvement of AR PR slices," Doc. JVT-T202, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, 20th Meeting, Klagenfurt, Austria, July 2006.