

SNR Scalability by Transform Coefficient Refinement for Block-Based Video Coding

Till Halbach[‡] and Thomas R. Fischer[†]

[‡]Department of Telecommunications
Norwegian University of Science and Technology (NTNU)
Trondheim, Norway

[†]School of Electrical Engineering and Computer Science
Washington State University (WSU) Pullman
WA, USA

ABSTRACT

Various techniques for SNR scalability in hybrid block-based video coding exist in the literature and in different standards. A new approach based on transform coefficient refinement is proposed in this article. The coefficient difference is computed after quantization, subsequently entropy-encoded, and transmitted for reconstruction of a high-quality layer at the decoder. The approach achieves in most cases only a moderate increase in bit rate as compared to other schemes. The bit rate of our two-layer framework converges towards the rate of a single-layer system as the quality difference between the two layers increase. The gains come at the cost of increased computational complexity and memory requirements.

Keywords: SNR/quality scalability, block-based hybrid video coding, bit stream layering, H.26L

1. INTRODUCTION

Currently, new video compression algorithms are being standardized by the Joint Video Team, a joint effort of ITU-T's Video Coding Experts Group and ISO/IEC's Moving Pictures Experts Group. The project will eventually evolve as two new – and identical – standards, Recommendation H.264 and International Standard 11172-10, also known as MPEG-4 Advanced Video Coding. The standardization process is more commonly addressed by its project name H.26L and has at the moment the status Draft International Standard. It is expected to be ratified by summer 2003.

H.26L is, among a number of other applications, being developed for broadcasting and streaming frameworks. On a client-server basis, these applications offer their services usually at different levels, i.e. with different image qualities for different clients. However, features like scalability by spatial resolution or quality have so far not been standardized in H.26L. Considering the latter scheme, this article proposes a new technique which is suitable for all hybrid block-based video codecs.

The article is structured as follows. First, we briefly introduce H.26L and explain its basic functionality. We then give an overview of different approaches to SNR scalability in previous research and standards, followed by the proposal for a new technique based on coefficient refinement. After that, we discuss implementation issues in detail and present simulation results. The article closes with concluding remarks.

2. DESCRIPTION OF H.26L

H.26L is very similar to prior digital video compression standards like H.26x and MPEG-x. It is a block-based hybrid coding scheme, i.e. a spatial or temporal prediction of the current signal is computed and subtracted from the original. Then, the prediction error is transformed, quantized and entropy-encoded. H.26L consists like its predecessors of several sets of algorithms, also called profiles, of which one has to be chosen for a specific application. We will concentrate in the following on the basic operation mode, which is specified by the Baseline profile. There, the transform is a DCT-approximating 4×4 integer transform. The transform coefficients are then processed by uniform scalar quantization and thereafter encoded by context-adaptive variable-length coding, resulting in a binary code stream. The choice among

At the time of writing, Till Halbach was on sabbatical leave at WSU in Pullman. The authors can be contacted by *halbach@tele.ntnu.no* and *fischer@eecs.wsu.edu*.

the 52 different available quantizers is made by passing a quantization parameter (QP) to the encoding system. The QP specifies the quantizer's step size and influences hereby both the encoder's output rate as well as the quality/ $PSNR$ of the reconstructed sequence. An increase of QP by one leads to an approximately 12% larger step size. To reduce complexity requirements, H.26L works on so-called macroblocks (MBs) which – assuming a YCbCr color space and 4:2:0 chrominance (chroma) subsampling – consist of one 16×16 luminance (luma) block and two 8×8 blocks for the color components. A group of MBs forms in turn a slice, the smallest independently decodable data unit.

The encoder has to determine a prediction mode for each MB with regard to an error criterion. This is referred to as mode decision and done independently for the luma and chroma signals with the exception that, if a luma MB is INTRA-coded, then also the predictions for both corresponding chroma MBs have to be intra-frame predictions. In order to adapt closely to the input signal, the codec follows a sub-MB/block concept. In INTRA (I) coding, a luma MB can either be predicted as a whole, 16×16 -pixel block, or is split into 16 blocks of size 4×4 . INTER (P) coding employs a hierarchical split where a luma MB can be cut into halves by dividing its dimensions horizontally, vertically, or both. The same scheme may in turn be applied to the resulting one to four 8×8 blocks. An H.26L system can hereby adapt to the detail of the input image and capture motion across frames.

In contrast to e.g. H.263, standardization of scalability by quality has not been considered in H.26L so far. However, SNR scalability is a very important feature which makes the system very flexible for applications in broadcasting and streaming, where different channel bandwidths are available with several receivers. It is further of advantage in error-prone environments for graceful degradation of the quality of the reconstructed pictures in case of transmission errors.

3. SNR SCALABILITY

SNR-scalable systems offer multiple representations of the visual data at different quality levels. This is achieved by several layers, each of which represents the original signal with a certain quality. The layer of lowest quality is the base layer. Higher-quality representations are accomplished by combining the data of a low-quality layer with data of one or several enhancement/refinement layers. A major drawback of block-based DCT systems which offer SNR scalability is the significant increase in bit rate as compared to a single-layer representation which represents the data at the highest quality of a multiple-layer framework. Hence, generally speaking, any framework which offers multiple-layer representations trades off bandwidth and SNR.

In the progressive operation mode of the JPEG standard and H.263 Annex O, SNR scalability is achieved by several coding passes. As a first pass/scan, coding of the original sequence at a low or moderate compression ratio leads to base layer data. This data is then reconstructed and subtracted from the original signal, hereby forming the coding error of the first pass. The error is in turn coded and sent to the receiver as refinement information. Theoretically, there can be an arbitrary but finite number of layers. Pennebaker and Mitchell report a 33% increase of the JPEG standard's bit rate with progressive coding as compared to a single-pass compression with comparable image quality.¹ Progressive coding in the JPEG standard is done by either spectral selection of DCT coefficients or by their successive approximation (bit plane coding), or a combination of both. Considering a layered H.263 Annex O video codec, an increase in bit rate of approximately 29% has to be expected to reach the quality of a single-scan coding scheme.² Here, scalability is achieved by coding the difference between original and reconstructed base layer signal, formed in the spatial domain.

A different scheme is standardized in MPEG-2 (in the SNR profile).³ The transform coefficients of the base layer are quantized with a high QP and form the base layer. The enhancement layer is derived by quantizing the base layer's quantization error with a low QP . The enhancement quantization error may or may not be fed back into the decoder loop of the encoder. With feed-back, a drift problem occurs because the information of the high-quality layer is used as a reference for prediction of the base layer. This approach is not suitable in error-prone environments as the decoder may be forced to use only the base layer signal for prediction in case the enhancement layer was lost under transmission.

In the following, we propose a new scheme which is suitable for SNR scalability in H.26L. To keep both the system simple and the amount of computation for simulations reasonable, we will focus on the instance of one base layer (BL) and one enhancement layer (EL).

4. TRANSFORM COEFFICIENT REFINEMENT

Due to the energy compaction property of most signal transforms in video coding, one would expect that schemes which form the coding error in the frequency domain are more efficient than schemes which compute the error in the spatial domain. Therefore, we compute the difference signal in the frequency domain *and* – in contrast to e.g. MPEG-2 – on the quantized transform coefficients. In other words, the transform coefficients of the second pass are predicted by the coefficients of the base layer, and the difference is then losslessly entropy-encoded. The efficacy of the approach is therefore dependent on the performance of the entropy-encoding module and the statistics of the scan of coefficient differences. If available, the refinement information is added to the transform coefficients on the decoder side for high-quality (HQ) image reconstruction. The approach of coding the transformed and quantized prediction error (e) difference $Q\{T\{e_{\text{HQ}}\}\} - Q\{T\{e_{\text{BL}}\}\}$ is more efficient than coding the reconstruction error $x_{\text{BL}} - \hat{x}_{\text{BL}}$ with regard to original signal x and reconstructed signal \hat{x} because of the high similarity of both signals, especially for large $\Delta QP = QP_{\text{BL}} - QP_{\text{HQ}}$. Additionally, side information like prediction mode and motion vector data is re-used with only insignificant quality degradation for the layer for which the mode decision has not been optimized. We stress the fact that, with the proposed two-layer technique, the high-quality layer reconstructed by QP_{HQ} is identical with a single layer which has been reconstructed setting $QP = QP_{\text{HQ}}$.

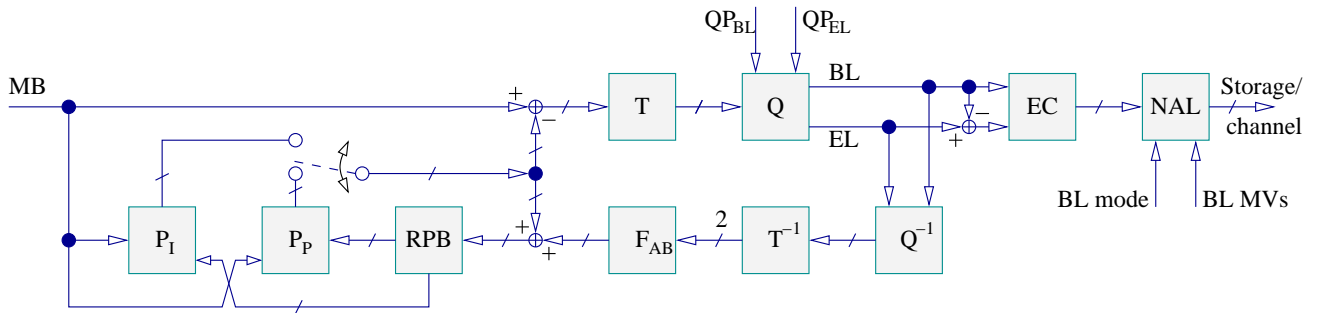


Figure 1. An H.26L-conform encoder modified for SNR scalability. Double internal nodes, marked e.g. by †, denote operation for all layers. Acronyms: P_I : INTRA prediction; P_P : INTER prediction; RPB: Reference picture buffer; F_{AB} : Anti-blocking filter; T (T^{-1}): (Inverse) transform; Q (Q^{-1}): (Inverse) quantization; EC: Entropy coding; NAL: Network abstraction layer; MVs: Motion vectors

The EL requires its own reference picture buffer and that prediction, transform and quantization be performed independently of the BL. However, it is mandatory for both layers to operate with the same prediction mode, i.e. macroblock type, since the two sets of transform coefficients are subtracted from each other. This means in particular that the time-consuming motion estimation is only done for one layer. Sharing side information like prediction mode and motion vectors is further the key to an efficient scalable solution which gives only a small bit rate increase. For streaming and videoconferencing systems, the encoder should optimize the mode decision with regard to the high-quality layer because it expects the application to receive ideally both code streams. For broadcasting applications and error-prone environments, mode decision should be chosen to be optimized with respect to the BL to provide a best possible image quality (quality of service) under erroneous transmission. The quality of the layer for which mode decision has not been optimized will degrade somewhat (typically around 1 dB in $PSNR$) when coding with the same QP but using a different layer to determine the optimal coding mode (INTRA /INTER, block size, reference picture). Subsequently, we have chosen the high-quality layer as basis for mode decision.

Due to the MB approach of H.26L, the EL does not necessarily have to cover the whole spatial plane but can be computed for one or several specific areas within, hereby allowing for region-of-interest (ROI) coding. These areas have to begin and end at MB boundaries because of the largest prediction mode which is of size 16×16 pixels both with INTRA and INTER coding. We propose the use of a 1-bit ROI array/mask, consisting of ones to signal ROI MBs and of zeros elsewhere, resulting in 99 data bits per frame for QCIF images, which have to be conveyed to the decoder. The ROI mask can, but does not have to, be updated for every video frame. This concept is very flexible as it allows for transmission of several areas of differently increased quality. However, a QP or ΔQP has to be signaled for each ROI. As the increase of overhead means a decrease of coding efficiency, we limit our implementation in the following to one ROI, i.e. the whole frame, with one fixed QP . An example for a high-quality layer which does not cover the whole picture is shown in Fig. 2.

We implemented the proposed scheme in both encoder and decoder of the Joint Video Team's reference software Joint



Figure 2. The foreground marked by the rectangle is coded by $QP_{\text{HQ}} = 20$, the background is coded with $QP_{\text{BL}} = 40$.

Model JM-4.0d. The corresponding preliminary standard document is the Joint Final Committee Draft.⁴ The software is operated in the Baseline profile mode which includes the features I/P slices, in-loop deblocking filter, 1/4-sample motion estimation/compensation, context-adaptive variable-length coding (CAVLC), and a YCbCr color space with 4:2:0 chroma subsampling. We set frame skip and number of reference frames to one and further the size of one slice equal to one sequence frame. As optimization criterion for low-complexity mode decision, we do Lagrange functional minimization with regard to the summed absolute transformed difference (*SATD*):

$$J = SATD + \lambda \cdot R, \quad (1)$$

with a rate constraint R and the optimization parameter λ . *SATD* for a 4×4 block is defined as

$$SATD = \frac{1}{2} \sum_{i,j=0}^3 |T_H\{D(i,j)\}|. \quad (2)$$

$T_H\{\cdot\}$ is the 2-D Hadamard transform, and the definition of the prediction error is

$$D(i,j) = X_{\text{org}}(i,j) - X_{\text{pred}}(i,j), \quad (3)$$

with the original and predicted samples X , at position pixel i in line j . The minimization is done for all intra- and inter-frame MB coding modes (the latter one implying all reference frames).

The encoder outputs, in addition to the BL bit stream, a separate EL stream which consists of a so-called coded block pattern (*CBP*) for each MB and the entropy-encoded transform coefficient differences for all color components, and both DC and AC coefficients. The *CBP* contains information about which of the scans of the six 8×8 blocks of a MB (four luma blocks and two chroma blocks) include non-zero coefficients. If an 8×8 block is marked as having only zero coefficients, no further data is conveyed for that block. We note that the *CBP* of the EL contains information about the scan *differences* of low- and high-quality layer. As H.26L works on a slice basis, we group also the MBs of the EL into groups and precede them with a unique start code, hereby adding 3 bytes to every slice. The EL may hence be transmitted packet-wise with one slice per packet, and the system is therefore appropriate for use in error-prone environments.

5. TESTING, RESULTS, AND DISCUSSION

We computed the bit rate increase for three different image sequences, both I and P slices/frames, and for various ΔQPs . The QP increment of five has been chosen because the bit stream sizes of BL and high-quality layer are approximately equal for $\Delta QP = 5$. The test sequences (and their resolutions) are high-motion sequence *Foreman* (QCIF), low-motion sequence *Container* (QCIF), and head-and-shoulder video *Salesman* (CIF), all with a frame rate of 30 Hz. All results have been averaged over 150 frames.

The values in Tab. 1 show two interesting tendencies: First, the bit rate increase varies significantly from quite large for small ΔQPs (here: two) to very small for large ΔQPs (e.g. 20). This can be explained by the fact that almost equal QPs

produce similar transform coefficient scans and an EL which consists of many one's and series of zeros within. Both are not efficiently entropy-encodable in a rate distortion sense. The worse the quality of the BL, the more of its coefficients are set to zero, and the more will the total bit rate converge towards the single-layer rate. Second, the rate increase gradient is more distinct for P frames than for I frames. The explanation is – similar to above – that the coefficient scans of P frames, coded with different QP s, are less correlated than the scans of I frames due to a more accurate inter-frame than intra-frame prediction. All results are consistent with these of the other sequences. We can conclude that our proposed scheme offers less increase in bit rate than the aforementioned systems for ΔQP s of five or greater.

QP_{HQ}	Frame type	QP_{BL}						
		20	22	25	30	32	35	40
20	I	0	35.6	37.7	27.1			9.8
20	P	0	44.7	30.8	15.9			4.2
25	I			0	32.2		20.9	18.7
25	P			0	30.0		16.5	10.4
30	I				0	33.8	30.0	18.7
30	P				0	52.2	40.2	25.2
35	I						0	29.7
35	P						0	51.8

Table 1. Average rate increase in % of the two-layer scheme as compared to the rate of a single layer. Sequence: *Foreman*; Coding mode optimization for high-quality layer. The results for P frames include an initial I frame.

We further computed the rate distortion (RD) curves with different QP s, as shown in Fig. 3. The five points of the curve labeled $\Delta QP = 0$ correspond – in order of decreasing SNR – to the single-layer quantization parameter QP ranging from 10 to 30 with an increment of five. We see that the rate increase varies strongly depending on ΔQP at high rates (> 500 Kbits/s) and less at low bit rates (< 200 Kbits/s). Generally, the smaller ΔQP , the higher the ΔR . The aforementioned observation that the total bit rate of base and high-quality layer converges with increasing ΔQP to the rate of a single-layer framework at high quality can again be made. To strengthen this observation, we computed the rate increase also for very high (≤ 30) ΔQP s, which is only possible at high rates due to the facts that the range of scalar quantizers and hereby QP is limited from 0 to 51. The bit rate increase is almost 0% when the sequence is reconstructed by means of a BL coded at $QP = 50$ and an HQ layer coded with $QP = 10$. Again, all results are consistent with these of the other sequences.

6. CONCLUSIONS

We have proposed the concept of transform coefficient refinement and layered bit streams for features like SNR scalability and ROI, and showed its application by inclusion of the new scheme in the video compression standard H.26L. The results show that the new scheme performs with an increase in bit rate comparable to the other approaches as mentioned in Sec. 3 when the difference between the two used QP s is approximately five and the total bit rate is larger than 200 Kbits/s.. A smaller ΔQP means a somewhat higher bit rate, up to about $\Delta R = 50\%$. Our scheme outperforms all other techniques in a rate distortion sense when ΔQP is sufficiently large, i.e. at least 5 at high and 10 at low rates, even without separate RD optimization in each layer as employed by Gallant and Kossentini.² We can hence conclude that the proposed scheme is an excellent candidate in error-prone environments and for systems where SNR scalability is highly desirable. The concept can further be extended to several layers; however, the maximum number of layers may be limited by memory requirements or restrictions with regard to computational complexity and latency.

ACKNOWLEDGMENTS

The authors wish to thank the Norwegian Research Council and the Department of Telecommunications for support and additional funding.

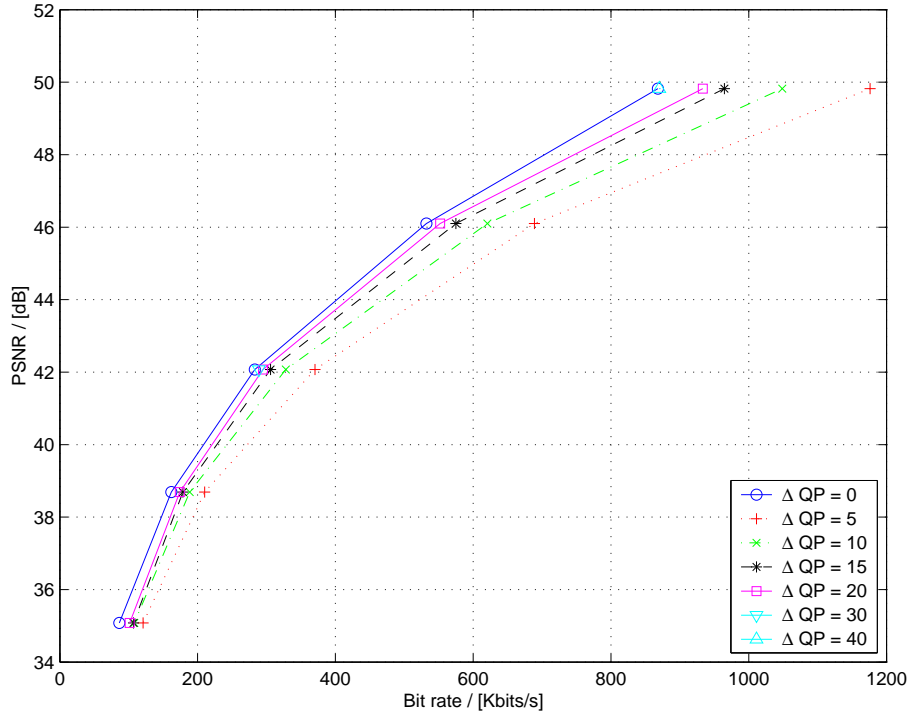


Figure 3. RD comparison. $\Delta QP = 0$ is the curve of the single-layer scheme. $\Delta QP > 0$ denotes the two-layer technique as explained with $\Delta QP = QP_{BL} - QP_{HQ}$.

REFERENCES

1. W. P. Pennebaker and J. L. Mitchell, *JPEG Still Image Data Compression Standard*, Van Norstrand Reinhold, New York (NY, USA), 1992.
2. M. Gallant and F. Kossentini, "Efficient scalable DCT-based video coding at low bit rates," in *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, (Kobe, Japan), Oct. 1999.
3. K. R. Rao and J. J. Hwang, *Techniques and Standards for Image, Video, and Audio Coding*, Prentice Hall, New Jersey (USA), 1996.
4. T. Wiegand, "Joint Final Committee Draft (JFCD) of joint video specification ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC," Tech. Rep. D157, ITU-T VCEG | ISO/IEC MPEG (JVT), Aug. 2002.