Title

Compression Efficiency and Delay Tradeoffs for Hierarchical B-Pictures and Pulsed-Quality Frames

Permalink

https://escholarship.org/uc/item/3bb5z9tm

Journal

IEEE Transactions on Image Processing, 16(7)

ISSN

1057-7149

Authors

Leontaris, A.
Cosman, P. C.

Publication Date

2007-07-01

DOI

10.1109/TIP.2007.896681

Peer reviewed

Compression Efficiency and Delay Tradeoffs for Hierarchical B-Pictures and Pulsed-Quality Frames

Athanasios Leontaris, *Member, IEEE*, and Pamela C. Cosman, *Senior Member, IEEE*

***Abstract*—Real-time video applications require tight bounds on end-to-end delay. Hierarchical bidirectional prediction requires buffering frames in the encoder input buffer, thereby contributing to encoder input delay. Long-term frame prediction with pulsed quality requires buffering at the encoder output, increasing the output buffer delay. Both hierarchical B-pictures and pulsed-quality coders involve uneven bit-rate allocation. Both the encoder and decoder buffering requirements depend on the rate allocation. We derive an efficient rate allocation for hierarchical B-pictures using the power spectral density of a wide-sense stationary process. In addition, we discuss important aspects of hierarchical predictive coding, such as the effect of the temporal prediction distance and delay tradeoffs for prediction branch truncation. Finally, we investigate experimentally the tradeoff between delay and compression efficiency.**

***Index Terms*—Bidirectional prediction, end-to-end delay, H.264/AVC, hierarchical B-pictures, motion-compensated temporal filtering (MCTF), motion compensation accuracy, video compression.**

I. Introduction

Constraining delay is critical for real-time communication and live event broadcast. Live television broadcast should have a delay of no more than one second in many cases. Interactive video-phone communication should have a maximum end-to-end delay of no more than 300 ms. For traditional predictive coding techniques where the current frame is predicted from the exactly previous frame, the end-to-end delay is low. A frame is captured, encoded in real-time, briefly buffered, and then transmitted. After brief buffering, the decoder decodes the bits and displays.

Throughout this work, the term *rate allocation* refers to the bit distribution on a frame basis. The allocation of this bit budget to individual macroblocks (MB) within a frame is denoted *rate control*. We note that rate control is possible not only by varying the quantization parameter (QP), but also through appropriate selection of coding modes, of the λ parameter of the lagrangian optimization, and the rounding of DCT coefficients. We also use the terms frame and picture interchangeably. Assuming that content is largely stationary and that we operate under very tight delay constraints, then rate allocation is straightforward: every frame receives the same number of bits so that a total bit rate constraint is satisfied. Compression efficiency can be improved either by increasing the buffering delay (bit rate allocated to each frame can vary) or when more flexible motion-compensated prediction (MCP) structures are used. These include prediction structures that use additional reference frames, as well as structures that use frames from both the past and the future.

By filtering across frames or by using bidirectional prediction, compression performance improves because the temporal correlation among several neighboring frames is better exploited, but additional delay is incurred. An example is motion-compensated temporal filtering (MCTF). Tradeoffs of delay and compression in MCTF video codecs were investigated in [2]. In that work, delay was reduced by selectively removing the update step. Recently, the update step was removed from the working draft of the Scalable Video Coding extension to H.264/AVC [3]. The end-to-end delay tradeoff for MCTF was studied in [4]. However, delay is an issue for hierarchical bipredictive structures, as well. The delay in the hierarchical case depends on the size of the group of pictures (GOP)[1] and cannot be reduced by removing update steps while keeping the GOP size intact. However, as we will show in this work, it can be reduced through the removal of motion-compensated prediction branches.

One can also have increased delay when using a single-direction (forward) prediction scheme. The codec proposed in [7] employs two reference frames, one short-term (ST) and one long-term (LT). The LT frame is afforded extra bits; it is of high (pulsed) quality. The rest of the frames are starved to achieve the rate constraint. At a given constant transmission bit rate, the LT frames will take longer to transmit, introducing delay. Compression efficiency was improved for certain image sequences, but delay was not studied in that work.

The studies in [2] and [4] did not take into account the effect of the encoder output and the decoder input buffering requirements which are nontrivial. Here, we model both delays by means of a comprehensive rate control scheme that is applied for the first time on B-coded pictures, in addition to P-coded pictures. In this work, we study the delay for LT pictures with

Manuscript received June 6, 2006; revised February 12, 2007. Part of this work appears in the 2006 IEEE International Conference on Image Processing. This work was supported in part by the Center for Wireless Communications at the University of California, San Diego; in part by the Office of Naval Research; and in part by the University of California Discovery Grant program of the State of California. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Fernando M. B. Pereira.

A. Leontaris is with Dolby Laboratories, Inc., Burbank, CA 91505-5300 USA (e-mail: aleon@dolby.com).

P. C. Cosman is with the Information Coding Laboratory, Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093-0407 USA (e-mail: pcosman@code.ucsd.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2007.896681

[1] In our work, we follow the usage of [5] and [6] in taking the term GOP size to mean the maximum number of frames that need to be buffered at the input (including the current frame) in order to encode the current frame. Furthermore, in this work, we concentrate on B-coded pictures that involve one frame from the future of the current frame and the other from its past. Generalized B-pictures are not within the scope of this work.

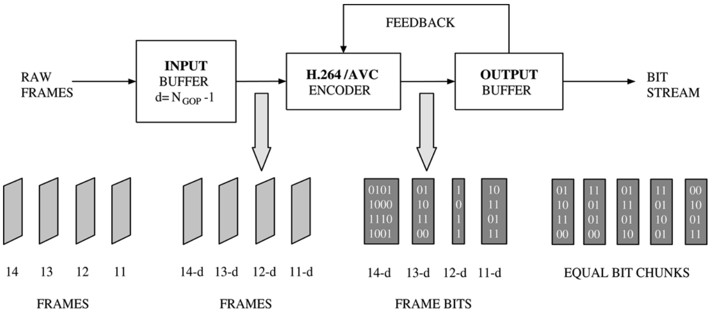


Fig. 1. Encoder input buffer and output buffer introduce delay.

pulsed quality, as well as for hierarchical B-coded pictures for varying GOP size. MCTF structures are not evaluated as they were found to be in most cases inferior to hierarchical B-coded pictures [8].

A key element of a delay-constrained video encoder is the rate control scheme. One of the early rate control algorithms was the Test Model 5 for MPEG-2 [9]. A complexity measure is calculated for each image and is then used to allocate the rate among the images. A similar approach can be found in [10]. An evolution of TM5 that replaced the block variance with the block sum of absolute differences (SAD) was proposed in [11] and further improved in [12]. However, an efficient algorithm that used a quadratic rate distortion model was proposed in [13] and formed the basis of the rate control that was adopted for the JM H.264/AVC reference software and whose detailed description can be found in [14] and [15]. Rate-control based on [14] was adopted for this work and is used in all codecs to ensure that the delay budget is enforced so that no buffer overflows occur.

Even when perfect rate control is possible, i.e., each frame receives exactly its preallocated number of bits, extra buffering delay at the encoder output and decoder input is incurred when the bit rate is distributed unevenly among the frames. We investigate the effect of uneven bit rate distribution on delay. It is noteworthy that bit rate allocation affects both compression efficiency as well as buffering delay. We are interested in trading off compression efficiency for less delay. This is possible if one has access to a sufficiently accurate rate-distortion model. Given constraints on bit rate and buffering delay, such a model can yield an efficient rate allocation within a GOP.

To obtain a model for hierarchical prediction, we need to account for the temporal prediction distance. In [16], a rate-distortion model was presented that modeled the 2-D video signal as a wide-sense stationary process. The rate and distortion were calculated as functions of the power spectral density of the prediction error. This model introduced the concept of motion-compensation accuracy, which was investigated in depth in [17]. Although in [16], motion compensation accuracy represented the level of fractional-pel accurate MCP, in this work, we intend to use this accuracy to also model the temporal prediction distance. Intuitively, as we attempt to predict from frames that are farther away, the predictor becomes less accurate.

The paper is organized as follows. In Section II, we define the end-to-end delay and describe several motion-compensated prediction structures. We calculate some useful bounds for delay constraints in the case of uneven bit rate allocation in Section IV. The motivation and the fundamental challenges of deriving a rate allocation scheme for hierarchical B-pictures, as well as the resulting estimate, are presented in Section V. The adopted rate allocation and control schemes for each codec are discussed in Sections VI and VII. Experimental results and conclusions follow in Section VIII. Finally, the paper is concluded in Section IX.

II. End-to-End Delay

End-to-end delay involves delay at the source encoder, channel encoder, channel decoder, and source decoder, as well as transmission and propagation delay. We assume a propagation delay of zero, and we assume a lossless channel, so we do not include any channel coding. We further ignore the actual computation time at the encoder and decoder, limiting our scope to the buffers at the source encoder, shown in Fig. 1, the transmission delay, and the buffers at the source decoder.

The first delay is at the encoder input buffer, and this delay depends on the motion-compensation structure used, and varies in increments of whole frame durations. The encoder converts the frame into a bit stream instantaneously and then starts writing the bits to the encoder output buffer at a constant rate. If frame i is encoded with b_i bits, then it is written into the buffer immediately at time instant $i \times t_{\text{fr}}$, where t_{fr} is the display time duration of a frame, which amounts to 33 ms for 30 frames per second. The above encoder output buffer is a "leaky bucket" [18], [19]: it is continuously drained at the constant average source coding bit rate r bits per second. This translates to $r/30$ bits being drained in the course of t_{fr} seconds. We assume constant bit rate (CBR) transmission. We note that the rate $30 \times b_i$ may be more or less than the average source coding rate r.

In this work, we strive to achieve an accurate target bit rate for each frame. However, CBR transmission can also be possible for time units larger than a frame. In those cases, the bit rate is considered constant on a time-unit basis and the constraint for equal-sized frames is removed. The encoder is free to allocate the rate within the frames of the time unit so as to optimize some (quality) criterion, while making sure that the unit as a whole adheres to the CBR target rate.

The encoder output buffer determines how tightly the rate allocation and rate control must operate. With a constant source coding rate of r bits per second, each frame could have the same exact $r/30$ bits per frame, and then the output buffer could be only $r/30$ bits long, but with a short output buffer, the encoder could not respond to a scene cut or to high motion by using

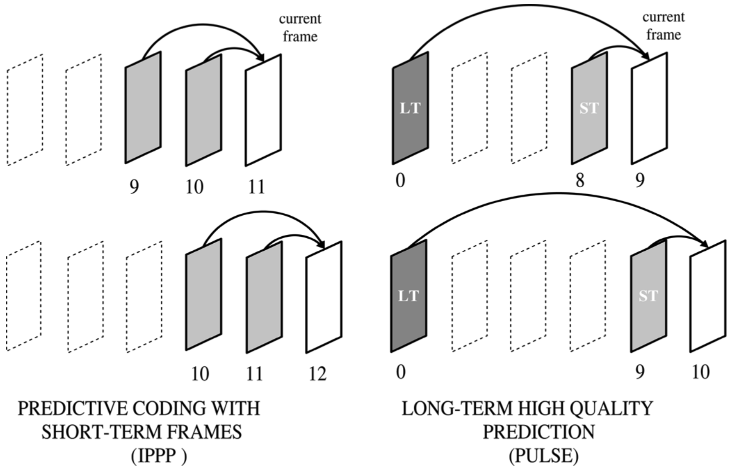

current frame
9
10
11
10
11
12
PREDICTIVE CODING WITH SHORT-TERM FRAMES (IPPP)
current frame
LT
ST
0
8
9
LT
ST
0
9
10
LONG-TERM HIGH QUALITY PREDICTION (PULSE)

Fig. 2. IPPPP and the PULSE motion-compensated structures. The arrows denote motion-compensated prediction.

more bits, and by giving fewer bits to static scenes. Allowing the encoder output buffer to be larger leads to higher video quality.

We will require our rate control to live within the buffer without any frame skipping and without producing overflows or underflows. Since every frame is written into the buffer instantaneously, the rate control must ensure that the length in bits of any single frame is no longer than the buffer size, and indeed, is no longer than the space remaining in the buffer at any particular time.

The decoder is a mirror image of Fig. 1. Bits are buffered in a decoder input buffer, which is the same size as the encoder output buffer, as discussed in [18]–[20]. According to [18] and [19], the encoder buffer fullness and the decoder buffer fullness are always complementary to each other and have a constant sum equal to the max size of each buffer. Decoded frames are buffered at the output prior to display. In our model, which excludes delays from computation, the source coding end-to-end delay D_{e2e} depends on the four buffers and can be written as

$$D_{e2e} = D_{\rm enc}^{\rm in} + D_{\rm enc}^{\rm out} + D_{\rm dec}^{\rm in} + D_{\rm dec}^{\rm out} \tag{1}$$

where subscripts indicate encoder or decoder, and superscripts indicate input or output buffers. Assuming that the size of the encoder output buffer and the decoder input buffer is B, we can write the associated delay as D_B and obtain: $D_{\rm enc}^{\rm out} + D_{\rm dec}^{\rm in} = D_B$.

We investigate five types of encoders: predictive IPPPP coding (IPPPP), long-term prediction with pulsed quality (PULSE), hierarchical B-pictures (HIER), a HIER codec with truncated prediction branches (TRUNC) that will be discussed in detail in Section III, and IBBBP where all B-coded pictures are disposable and use only I-and P-coded pictures as references. The codecs are now described in detail.

The IPPPP codec, shown in Fig. 2, is based on the Joint Model (JM) 10.1 reference software of the H.264/AVC video coding standard [21], [22]. Frames are encoded predictively in an I-P-P-P-P structure. The IPPPP coding order can also be written as: I0 P1 P2 P3 P4. In this notation, the letter in each letter-number pair denotes the H.264/AVC slice type, and the number denotes the display order.

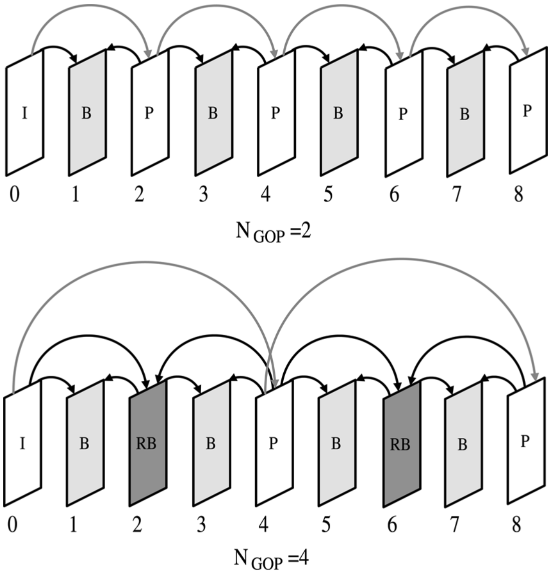


Fig. 3. Hierarchical bipredictive motion-compensated structures. The arrows denote motion-compensated prediction. The RB frames are B-coded pictures that can be used as *references*.

The PULSE codec, shown in Fig. 2, uses a short-term (ST) reference frame and a long-term (LT) reference frame for motion-compensated prediction as described in [7]. It is based on a modified version of the JM 10.1 reference software. The LT reference frame is periodically updated every U frames and is afforded more bits than the regular frames. Let $N_{\rm GOP}$ denote the number of frames in a GOP. Both for IPPPP, as well as for PULSE, we have $N_{\rm GOP} = 1$ and the coding order is the same: I0 P1 P2 P3 P4.

The HIER coder uses hierarchical picture motion-compensated prediction. These prediction structures, called hierarchical B-pictures, are composed of more than one temporal resolution level (a hierarchy). The simplest case is the well-known IBPBP prediction structure. Examples for $N_{\rm GOP} = 2$ (IBPBP) and $N_{\rm GOP} = 4$ are shown in Fig. 3. The coding orders for IBPBP and HIER with $N_{\rm GOP} = 4$ are I0 P2 b1 P4 b3 and I0 P4 B2 b1 b3, respectively. Small “b” denotes disposable B-coded pictures and capital “B” denotes B-coded pictures that are used as

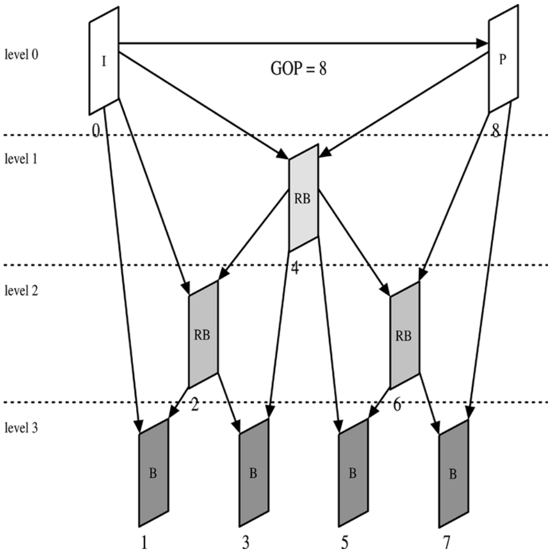


Fig. 4. Hierarchical B-Pictures for $N_{\rm GOP} = 8$.

prediction references. Fig. 4 illustrates a $N_{\rm GOP} = 8$ structure. HIER coders have by definition $N_{\rm GOP} > 1$. The number of hierarchical temporal levels is given by $\log_2 N_{\rm GOP} + 1$. The JM 10.1 reference software was used to code image sequences with HIER codecs as it already supports hierarchical B-coded pictures. Hierarchical structures benefit from prediction both from the "future" and the "past." This is particularly advantageous in cases of global motion and camera pan as shown in [23]. Note that the "closed loop" approach [8] was used for hierarchical prediction: B-coded pictures are predicted from the reconstructed reference frames and not the original ones as traditionally done in MCTF.

With hierarchical B-coded pictures, the encoder cannot begin to encode a frame until the entire GOP is available for processing. For GOP size equal to $N_{\rm GOP}$, the encoder begins processing one frame while $N_{\rm GOP} - 1$ frames are in the encoder input buffer. Thus, $D_{\rm enc}^{\rm in} = (N_{\rm GOP} - 1)t_{\rm fr}$, where $t_{\rm fr}$ is the display time duration of a frame.

For the IBBBP coder, we have a special case, where the notion of $N_{\rm GOP}$ changes from the encoder to the decoder. At the encoder side, the input delay is going to be equal to $(N_{\rm GOP}-1)t_{\rm fr}$. For the IBBBP case, the extra output delay at the decoder is constant and corresponds to the equivalent delay of a HIER coder with $N_{\rm GOP} = 2$. For example, in the case of HIER with $N_{\rm GOP} = 4$, after frame 4 has been decoded, the decoder has to decode frame 2 before it can decode frame 1 and, finally, frame 3. In contrast, in the IBBBP case, frame 1 (which is always the worst case scenario in terms of delay) can be decoded right after frame 4 has been decoded. This is true for *any* number of B-coded pictures used between the I-/P-coded pictures. The coding order in this case is I0 P4 b1 b2 b3.

In all our experiments, we encode at 15 or 30 frames per second, so we obtain $t_{\rm fr} = 66$ ms or $t_{\rm fr} = 33$ ms. Finally, the output/display decoder delay, for all codecs apart from IBBBP, is $D_{\rm dec}^{\rm out} = (\log_2 N_{\rm GOP})t_{\rm fr}$. For IBBBP, we write it as $D_{\rm dec}^{\rm out} = 2 \times t_{\rm fr}$. We, thus, rewrite (1) as

$$\begin{aligned} D_{e2e} &= D_{\rm enc}^{\rm in} + D_B + D_{\rm dec}^{\rm out} \\ &= D_B + (N_{\rm GOP} - 1)t_{\rm fr} + D_{\rm dec}^{\rm out}. \end{aligned} \tag{2}$$

If the rate allocation could achieve exactly $r/30$ bits per frame, then only a buffer of length $r/30$ would be needed at the encoder output or decoder input, and the above result shows that the delays for $N_{\rm GOP}$ equal to 1, 2, 4 would be 1, 3, and 6, respectively, times the frame duration of 33 ms. In all codecs we use a single I-coded picture at the beginning of the image sequence and it is coded with a constant QP that was selected as follows: the sequence was encoded once, disregarding delay constraints, for the given target rate. The average QP for the entire sequence is calculated and after incrementing it by two it is used to encode the first I-coded picture. Rate control is applied on the rest of the frames which are P- or B-coded pictures.

For the IPPPP and PULSE codecs, we derive from (2) the end-to-end delay as $D_{e2e} = D_B$. Recall that both IPPPP and PULSE have $N_{\rm GOP} = 1$. However, the buffer size B for PULSE will be larger to ensure good performance, since the high quality frames require more bits.

III. Structural Delay Tradeoff

So far, we have discussed the delay tradeoff with respect to encoder output buffering that is a direct result of the rate allocation scheme. However, from (1), we note that end-to-end delay is also a function of structural delay, which is nontrivial in hierarchical prediction systems. A study on reduced structural delay appeared in [4] for MCTF systems, and reduced structural delay has been integrated into hierarchical prediction structures. In this section we study the effect of *branch removal* from hierarchical B-picture coders.

To illustrate our examples, we will use the case where GOP size is set to 4. Such a prediction structure is illustrated on the left of Fig. 5. Frame 2 is predicted from frames 0 and 4, then frame 1 is predicted from frames 0 and 2, and frame 3 is predicted from frames 2 and 4. Assume now that the prediction of frame 2 from frame 4 is removed, as shown in the middle of Fig. 5. This brings down the structural delay by one half: the truncated GOP size 4 structure has $N_{\rm GOP} = 2$ instead of 4. The structure is similar to a GOP size 2 structure, shown on the right of Fig. 5. There are only two differences: a) it still has three hierarchical levels and allows more granular temporal scalability or network condition adaptability (frames 1, 2, and 3 can be dropped without affecting the reconstruction of frame 4) and b) frame 4, instead of being predicted from frame 2 as for GOP size 2, is predicted from frame 0. This means that compression performance will be worse than for a GOP size 2 structure, since the temporal prediction distance for frame 4 increases. Hence, for hierarchical B-pictures, the tradeoff of compression efficiency for delay effectively becomes a tradeoff of compression efficiency for increased temporal scalability and bitstream resilience and decreased delay.

The rate allocation scheme presented in Section V-C can be used to derive an efficient bit rate distribution in those cases where prediction branches are truncated. The bit rate distribution is adapted to reflect the removal of the branch and reflect single over double hypothesis. Indicative experimental results are provided in Section VIII.

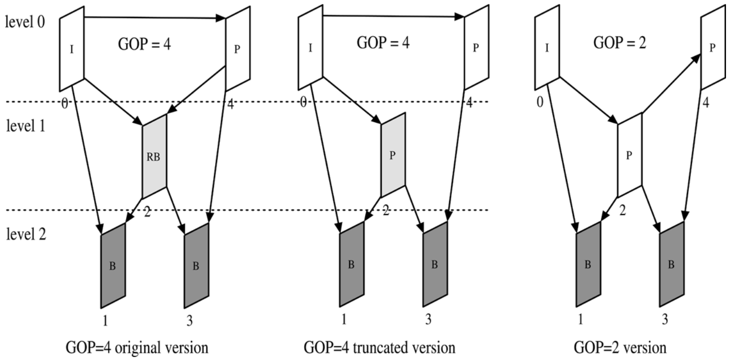


Fig. 5. (Left) Original hierarchical prediction structure for $N_{\text{GOP}} = 4$, (middle) its truncated version, and (right) the original hierarchical prediction structure for $N_{\text{GOP}} = 2$.

IV. Delay Calculation for Uneven Rate Allocation

In this section, we calculate the delay due to buffering at the encoder output. Both for pulsed quality dual frame video coding, as well as for hierarchical B-pictures coding, the problem of increased encoder output buffering is unavoidable when good compression efficiency is desired.

The following assumptions are made: a) bit rate is controlled by varying the QP and block coding mode to achieve the allocated target rate for this frame, and b) the rate allocation mechanism is accurate enough to ensure that within a block of K frames it allocates exactly $K \times R$ bits, where R is the constant bit rate. The block of K frames associated with the rate allocation periodicity is not necessarily the same as the GOP associated with the prediction structure periodicity. We will refer to the block of K frames as GOP-R. We take the length K of the GOP-R to be equal to N_{GOP} except for the case of the PULSE coder, for which we take K to be equal to U, the long term frame updating period.

Let x_i, where $0 \leq x_i \leq K \times R$, denote the rate allocated to frame i. Let $x_M = \max_{i \in [0, K-1]} x_i$ denote the maximum value of x_i. Let B denote the encoder output buffer length in bits. To avoid a buffer overflow during encoding, the necessary condition is

$$B \geq \max_{j \in [0, K-1]} \left[\sum_{i=0}^{j} \max(x_i - R, 0) \right]. \qquad (3)$$

The encoder can estimate the encoder output buffer length from the bit rate allocation. A useful and intuitive lower bound can be written as

$$B > \max(x_0, x_M, R). \qquad (4)$$

The buffer size should be larger than the first frame, the largest frame (in terms of bits), and the constant rate per frame R. We note that very often the first frame in a GOP-R is the frame that is allocated the highest single number of bits in the GOP-R.

In video coding systems, where the initial frame is pulsed and the later frames are starved (not necessarily equal in size), the ensuing delay depends on the rate allocated to the pulsed frame. Hence, the decision on the rate allocated to the pulsed frame trades off delay for compression performance. This result is valid both for hierarchical B-picture coding with GOP-R size $K = N_{\text{GOP}}$, as well as for pulsed quality dual frame coding with GOP-R size K equal to updating period $K = U$. It translates the delay constraint into a rate allocation constraint.

Allocating excessive rate to a frame, and starving the rest, can not only decrease performance, but can also dramatically increase delay due to buffering. In the next section (Section V), we seek to find a good theoretical rate allocation for a hierarchical structure. For pulsed quality, we derive experimentally an efficient rate allocation scheme later in this paper.

V. Proposed Framework for Rate Allocation

A. Motivation

The original rate allocation in the JM reference software video encoder uses a single QP for the entire frame. The QP value allocated to a bidirectionally predicted picture (B-coded pictures) is higher (coarse quantization) than the average of the QPs used to encode the two references. This arrangement guarantees high compression efficiency. However, rate allocation under tight delay constraints cannot use the same QP for the entire frame. To achieve a rate constraint, the QP is changed periodically within the picture. We can only set the bit rate for each frame. The per-block QP decisions seek to avoid buffer overflow and underflow and satisfy the target rate. We note that the original JM 10.1 rate control allowed per-block decisions only for P-slices. We extended this rate control algorithm to offer per-block decisions in cases of B-slices as well.

Our goal is to establish the bit rate allocation for different hierarchical levels with B-coded pictures, shown in Fig. 4. It is known that bidirectional prediction (multihypothesis prediction with two hypotheses) attenuates the prediction error energy by half, compared to uni-directional prediction [24]. Furthermore, we have to take into account the temporal distance from the reference frames. We found through experiments that the efficiency of the bidirectional prediction of a frame depends on the distance from its references. The QP allocation algorithm in the JM ignores this distance. However, the temporal distance determines the motion compensation accuracy and hence the resulting prediction error. Our goal is to model the influence of prediction temporal distance on compression efficiency.

We now discuss our main assumptions: a) frames within a temporal decomposition level have similar entropy, b) closed-loop coding, and c) high-rate operation.

We assume that the image sequence is correlated enough so that frames within the same temporal decomposition level have similar entropies and can be afforded the same number of bits. We seek a solution that does not depend on video content: fixed proportion of bits for each temporal decomposition level, divided equally among the frames of the level. Given the overall bit-budget and the proportions, it is straightforward to calculate the exact rates. The requirement for fixed ratios is a result of computational and delay constraints: the complexity and delay needed to optimize the rate allocation for each sequence are prohibitive.

We are primarily interested in rate allocation for closed-loop hierarchical B-pictures. Closed-loop refers to using as references the previously reconstructed versions of the frames. It always outperforms open-loop prediction. Optimal rate allocation for closed loop prediction is significantly harder. It is essentially a problem of dependent quantization [25]. Approaches for rate allocation in open loop MCTF based on temporal propagation of the error have been proposed and refined in [26]–[28] and [29]. Rate allocation for MCTF is relatively easy to obtain, since MCTF is an open-loop (the motion compensation references are the *original* frames) video encoder. Still, those approaches did not take into account the temporal distance between the frames and lack any delay constraints. They are primarily modeling the error attenuation due to temporal filtering and are not appropriate for this work since we cannot afford the delay and the computational complexity to analyze the signal and derive near-optimal rate allocation.

We assume operation at high rates. It was shown in [30] that closed-loop prediction at high rates does not alter the signal significantly. Hence, the effect of quantization error on prediction efficiency can be neglected for sufficiently fine quantization [31]. It was then suggested in [32] that using a closed-loop video coder with the optimal open-loop rate allocation performs close to the optimal closed-loop rate allocation. We will use the theory, originally developed in [16], and later extended to multihypothesis prediction in [24], to model rate-distortion behavior in hierarchical B-picture prediction. Open-loop compression efficiency for MCTF-like structures was investigated in [33] using the methodology of [16]. That approach, however, does not yield bit allocations on a frame basis and it also assumes a Karhunen–Loève transform which is not realistic in either traditional MCTF or closed-loop B-coded pictures.

The main reason we prefer the model of [16] over other approaches for bit allocation, is that results presented in [16] and [24] depend on the motion compensation accuracy. We propose that this accuracy is a direct function of the temporal distance between the reference and the predicted frame. A second reason is the modeling of multihypothesis prediction coding efficiency (B-coded pictures involve two hypotheses).

Last, we note that even though our simulations use the H.264/AVC video coding standard, we do not model the effect of the loop filter in the next paragraphs. The main reason is complexity, and the fact that, in the context of our work that prizes low delay and makes some sacrifices in compression efficiency, the additional gain by an accurate model is outside the scope of this work.

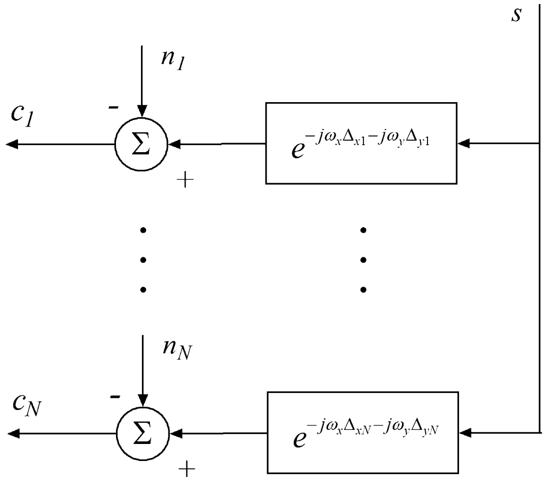

s
n_1
c_1
-
Σ
+
$e^{-j\omega_x\Delta_{x1}-j\omega_y\Delta_{y1}}$
n_N
c_N
-
Σ
+
$e^{-j\omega_x\Delta_{xN}-j\omega_y\Delta_{yN}}$

Fig. 6. Signal model for multihypothesis prediction from [24].

B. Theoretical Background

We will briefly now outline the rate-distortion (R-D) modeling scheme from [16] and [24].

ω_x	Horizontal frequency.
ω_y	Vertical frequency.
$\Phi_{ss}(\omega_x, \omega_y)$	Signal power spectrum of the input video signal.
$F(\omega_x, \omega_y)$	Frequency response of the "loop filter."
$P(\omega_x, \omega_y)$	2-D Fourier transform of the displacement error p.d.f.
$\Phi_{nni}(\omega_x, \omega_y)$	Power spectrum of residual noise component i that cannot be predicted by motion compensation.
$\Re(\cdot)$	Real part of a complex number.
D	Distortion resulting from encoding a signal with R bits per sample.

The original signal $s(x,y)$ is predicted by summing the convolutions of the spatial 2D convolution filters $f_i(x,y)$ with the hypotheses $c_i(x,y)$ (reference frames), where $i = 0, \cdots, N-1$, and N is the number of hypotheses used. The prediction error can then be written as

$$e(x,y) = s(x,y) - \sum_{i=0}^{N-1} f_i(x,y) * c_i(x,y). \qquad (5)$$

The signal model is depicted in Fig. 6, borrowed from [24]. The c_i are the hypotheses which are assumed to be versions of the original source signal s, corrupted with white noise n_i, and also shifted in 2-D by Δ_{xi} in the horizontal direction and Δ_{yi} in the vertical direction. The Δ_x and Δ_y are also modeled as random variables. In this work, we assume that the p.d.f. $P(\omega_x, \omega_y)$ of the displacement error Δ_x and Δ_y is a function of the temporal prediction distance. If the power spectral density of the prediction error $\Phi_{ee}(\omega_x, \omega_y)$ is known, then the error variance σ_e^2 is given from Parseval's relation as

$$\sigma_e^2 = \frac{1}{4\pi^2} \int_{-\pi f_{sx}}^{+\pi f_{sx}} \int_{-\pi f_{sy}}^{+\pi f_{sy}} \Phi_{ee}(\omega_x, \omega_y) d\omega_x d\omega_y \qquad (6)$$

where terms f_{sx} and f_{sy} are the spatial sampling frequencies in the horizontal and the vertical direction. The well-known rate distortion function for memoryless coding is

$$R(D) = \frac{1}{2}\log_2\left(\frac{\sigma_e^2}{D}\right) \tag{7}$$

in bits per sample (pixel). The power spectrum $\Phi_{\mathrm{ee}}(\omega_x, \omega_y)$ is calculated for N-hypothesis prediction in [24]. Since, in this work, we are studying single and double hypothesis prediction, we need the expressions for $N = 1$ and $N = 2$. For $N = 1$ hypotheses (P-coded pictures) (23) in [24] yields the following expression:

$$\begin{aligned}\frac{\Phi_{ee}(\omega_x,\omega_y)}{\Phi_{ss}(\omega_x,\omega_y)} = 1 &+ \left|F_{(1)}(\omega_x,\omega_y)\right|^2 \\ &- 2\Re\left\{F(\omega_x,\omega_y)P_{(1)}(\omega_x,\omega_y)\right\} \\ &+ \frac{\Phi_{nn1}(\omega_x,\omega_y)}{\Phi_{ss}(\omega_x,\omega_y)}\left|F_{(1)}(\omega_x,\omega_y)\right|^2.\end{aligned} \tag{8}$$

For $N = 2$ hypotheses (B-coded pictures), the power spectral density is given from (23) in [24] as

$$\begin{aligned}\frac{\Phi_{ee}(\omega_x,\omega_y)}{\Phi_{ss}(\omega_x,\omega_y)} = 1 &- 2\Re\left\{F_{(2)}\begin{bmatrix}P_{(2,1)}\\ P_{(2,2)}\end{bmatrix}\right\} \\ &+ F_{(2)}\begin{bmatrix}1+\alpha_1 & P_{(2,1)}P_{(2,2)}^* \\ P_{(2,2)}P_{(2,1)}^* & 1+\alpha_2\end{bmatrix}F_{(2)}^H.\end{aligned} \tag{9}$$

Terms α_i for each hypothesis are given in (22) in [24] as $\alpha_i = \Phi_{nni}(\omega_x, \omega_y)/\Phi_{ss}(\omega_x, \omega_y)$, where the power spectrum of the signal s is found in (19) in [17] as

$$\Phi_{ss}(\omega_x,\omega_y) = \frac{2\pi{\sigma_s}^2}{\omega_0^2}\left(1+\frac{\omega_x^2+\omega_y^2}{\omega_0^2}\right)^{-\frac{3}{2}} \tag{10}$$

where ${\sigma_s}^2$ is the variance of the original signal s. The noise power spectrum is $\Phi_{nn}(\omega_x, \omega_y) = \sigma_n^2$. Terms $P_{(1)}(\omega_x, \omega_y)$, $P_{(2,1)}(\omega_x, \omega_y)$, and $P_{(2,2)}(\omega_x, \omega_y)$ are critical and relate to motion compensation accuracy. The first is the displacement error p.d.f. for a P-coded picture, while the latter two correspond to each of the hypotheses in a B-coded picture. In [16] and [24], they are not parameterized with distance and are assumed i.i.d. with p.d.f.

$$p(\Delta_x,\Delta_y) = \frac{1}{2\pi\sigma_\Delta^2}e^{-\frac{\Delta_x^2+\Delta_y^2}{2\sigma_\Delta^2}}. \tag{11}$$

The Fourier transform of the above probability density function is calculated as

$$\begin{aligned}P(\omega_x,\omega_y) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p(\Delta_x,\Delta_y)e^{-j\omega_x\Delta_x}e^{-j\omega_y\Delta_y}d\omega_x d\omega_y \\ &= e^{-2\pi\sigma_\Delta^2(\omega_x^2+\omega_y^2)}.\end{aligned} \tag{12}$$

Finally, terms $F_{(1)}(\omega_x, \omega_y)$ and $F_{(2)}(\omega_x, \omega_y)$ from (8) and (9) represent the Fourier transform of the spatial filters $f(x, y)$ for single and double hypothesis. For $N = 1$ hypotheses (P-coded pictures), the Fourier transform is equal to $F_{(1)}(\omega_x, \omega_y) = 1$ (we assume no loop filtering) since $f_{(1)}(x, y) = \delta(x, y)$. For $N = 2$ hypotheses, it is set to $F_{(2)}(\omega_x, \omega_y) = [1/2 \quad 1/2]$ since $f_{(2)}(x, y) = [1/2\delta(x, y) \quad 1/2\delta(x, y)]$. The hypotheses are simply averaged.

We continue the derivation of the power spectral densities of (8) and (9). To simplify notation, we adopt $\Lambda = (\omega_x, \omega_y)$ from [16]. For the $N = 1$ hypothesis, we obtain the following expression:

$$\Phi_{ee}^{(1)}(\Lambda) = \Phi_{nn}(\Lambda) + 2\Phi_{ss}(\Lambda)\left(1 - P_{(1)}(\Lambda)\right). \tag{13}$$

For $N = 2$, we derive

$$\begin{aligned}\Phi_{ee}^{(2)}(\Lambda) = \Phi_{ss}(\Lambda)(1 + 0.5(1 &+ \frac{\alpha_1}{2} \\ &+ \frac{\alpha_2}{2} + P_{(2,1)}(\Lambda)P_{(2,2)}(\Lambda)) \\ &- P_{(2,1)}(\Lambda) - P_{(2,2)}(\Lambda)).\end{aligned} \tag{14}$$

We assume that both hypotheses are distorted equally yielding $\alpha_1 = \alpha_2 = \alpha = \sigma_n^2/\Phi_{ss}(\Lambda)$. Terms $P_{(2,1)}(\Lambda)$ and $P_{(2,2)}(\Lambda)$ represent the motion compensation accuracy for each of the two hypotheses. They are also assumed to be equal, so we define $P_{(2)}(\Lambda) = P_{(2,1)}(\Lambda) = P_{(2,2)}(\Lambda)$. Given the above two assumptions we finally obtain the following expression for the prediction error power spectral density:

$$\Phi_{ee}^{(2)}(\Lambda) = \frac{1}{2}\Phi_{nn}(\Lambda) + \frac{1}{2}\Phi_{ss}(\Lambda)\left(3 + P_{(2)}^2(\Lambda) - 4P_{(2)}(\Lambda)\right). \tag{15}$$

Comparing (13) with (15), we notice that the power spectral density of the residual error n is attenuated by $(1/2)$. This is an intuitive property of the selected averaging filter $F(\omega_x, \omega_y)$. It is obvious that if σ_n^2 is larger, then there is a larger benefit in going to two hypotheses. Another property that is easy to identify is that if $P_{(2)}(\Lambda) = P_{(1)}(\Lambda)$, then two hypotheses are always better than a single hypothesis since$\Phi_{ee}^{(1)}(\Lambda) - \Phi_{ee}^{(2)}(\Lambda) = 1/2[\Phi_{nn}(\Lambda) + \Phi_{ss}(\Lambda)(1 - P^2(\Lambda))] \geq 0$. Naturally, there can be cases where the previous inequality is not upheld. These are cases where $P_{(2)}(\Lambda) \neq P_{(1)}(\Lambda)$.

C. Proposed Estimate

We parameterize $P(\Delta_x, \Delta_y)$ with respect to the temporal prediction distance by setting $\sigma_\Delta = \sigma_\Delta(\Delta_t)$. In a hierarchical B-coded picture system with $L + 1$ temporal decomposition levels, the prediction distance Δ_t decreases as the temporal resolution increases. For $L = 1$, which corresponds to a GOP size of 2 and is the well-known IBPBPB structure, the distance Δ_t is 2 frames at level 0 and 1 frame at level 1. Let l denote the temporal decomposition level. We can write the prediction distance with respect to the level and the GOP size N_{GOP} as

$$\Delta_t(l) = 2^{-l}N_{\mathrm{GOP}}, l \in [0, L]. \tag{16}$$

We assume full binary temporal decomposition, hence $N_{\mathrm{GOP}} = 2^L$. Let us assume that $\sigma_\Delta(\Delta_t)$ is known, and that the optimal open-loop rate allocation is sufficiently close to the optimal closed-loop allocation. This allows us to find the optimal rate allocation with a single R-D curve for each of the

$L+1$ hierarchical levels.[2] The rate allocation is obtained by constraining the same distortion D across all $L+1$ curves.

The above technique relies on estimating $\sigma_\Delta(\Delta_t)$. We, therefore, encoded several sequences for varying frame rates where we constrained equal rate on a frame basis. Two of those sequences are shown in Fig. 13. We used the JM 10.1 H.264/AVC reference software to encode the sequences. Single-frame/hypothesis prediction was used. The rate control operated on groups of 11 macroblocks. The different temporal distances were obtained by varying the temporal subsampling ratio. The graphs present the distortion as mean squared error (MSE) versus bits per sample (pixel) and the curves correspond to different temporal distances. A frame rate of 30 fps leads to $\Delta_t = 1$, 15 fps has $\Delta_t = 2$, 7.5 fps leads to $\Delta_t = 4$, and so on. The figures show that the MSE σ_e^2 curves move to the right and to the top with an approximately logarithmically spaced rate. Hence, σ_e^2 seems to be a logarithmic function of the temporal distance Δ_t. We now need to establish the relationship between Δ_t and σ_Δ.

Replacing Φ_{ee} in (6) with the expression derived in (13), we obtain the theoretical performance for a single-hypothesis hybrid video encoding system with memoryless encoding of the prediction error. This is illustrated in Fig. 14. Term $\sigma_\Delta \omega_0$ is varied at equal spaces and produces approximately logarithmically spaced rate-distortion functions. This is important as R-D curves from our experimental investigation appear approximately logarithmically spaced, as well. We conclude that for fixed ω_0 the standard deviation of the motion compensation displacement error σ_Δ varies approximately *linearly* with the temporal prediction distance Δ_t. Therefore, we can estimate σ_Δ as

$$\tilde{\sigma}_\Delta(\Delta_t) = \alpha + \beta(\Delta_t - 1). \tag{17}$$

The next challenge is to estimate the constants α and β. The H.264/AVC reference software uses quarter-pel accurate motion compensation. From [24] we know that the value of $\tilde{\sigma}_\Delta(1) = \alpha$ is calculated to be approximately 0.0702 for quarter-pel MCP. For half-pel motion compensation it doubles to 0.1404, and again doubles to 0.2808 for integer-pel motion estimation. Parameter β was estimated as 0.108 by fitting the theoretical curves in Fig. 14 to experimental data, some of which were presented in Fig. 13. It is kept constant for all sequences. This value is a compromise for all the sequences tested. Naturally, if one were to optimize these values for each individual sequences given their content statistics, the performance could be further improved. We note that the above estimated parameters α and β are valid for the specific values of ω_0, σ_s, and σ_n we selected.

The final rate-distortion model is written as

$$R_l = \frac{1}{2}\log_2\left(\frac{\sigma_e^2(\Delta_t, N)}{D \times (1 + \epsilon \times (l+1))}\right) \tag{18}$$

where R_l is the bit rate allocated to a frame at hierarchical decomposition level l. Parameter N denotes the number of hypotheses used to predict that frame. Term σ_Δ is estimated by $\tilde{\sigma}_\Delta$ in (17) and is then plugged into $\Phi_{ee}^{(N)}(\omega_x, \omega_y)$ to yield $\sigma_e^2(\Delta_t, N)$. The motivation behind adding the term $\epsilon \times (l+1)$ to the denominator of (18) is that hybrid video coding is closed-loop and thus a case of dependent video coding. Frames at temporal level l are predicted from frames of level $l-1$ or less. However, these reference frames have already been quantized and the R-D curves of the current frame will have shifted as a result. The constant ϵ approximates this shift. The constant parameter ϵ was empirically set to a small value $\epsilon = 0.1$.

VI. Rate Allocation Algorithm

We now discuss how we allocate the bit rate to each frame for hierarchical B-coded pictures and pulsed long-term quality frames. The allocation of that rate to individual macroblocks is the job of the rate control scheme that follows in Section VII.

The bit rate allocation problem is stated as follows: given a target average bit rate of R_f bits per frame, we seek the rate that should be allocated to frames that belong to a specific hierarchical level. The obtained rate R_l values from (18) are used to establish bit rate ratios among temporal levels. Assuming for example $N_{\mathrm{GOP}} = 4$, we have three temporal levels: 0, 1, and 2. Level 0 contains the P-coded pictures for which $\Delta_t = 4$ and $N = 1$. Level 1 contains RB-coded pictures (RB in H.264/AVC notation are B-coded pictures that can be referenced during motion compensation) for which $\Delta_t = 2$ and $N = 2$. Finally, at level 2, we have disposable B-coded pictures (in H.264/AVC terminology it means that these frames cannot be used as references) for which the temporal distance is $\Delta_t = 1$ and $N = 2$. Using our algorithm we fix a common D, say an MSE of 20, that corresponds to a PSNR of 35.12 dB. We, thus, obtain R_0, R_1, and R_2 from (18). Given the R_i terms, we can derive the rate ratios between each level; we note that these do not yet represent the exact bits allocated to a frame in those levels.

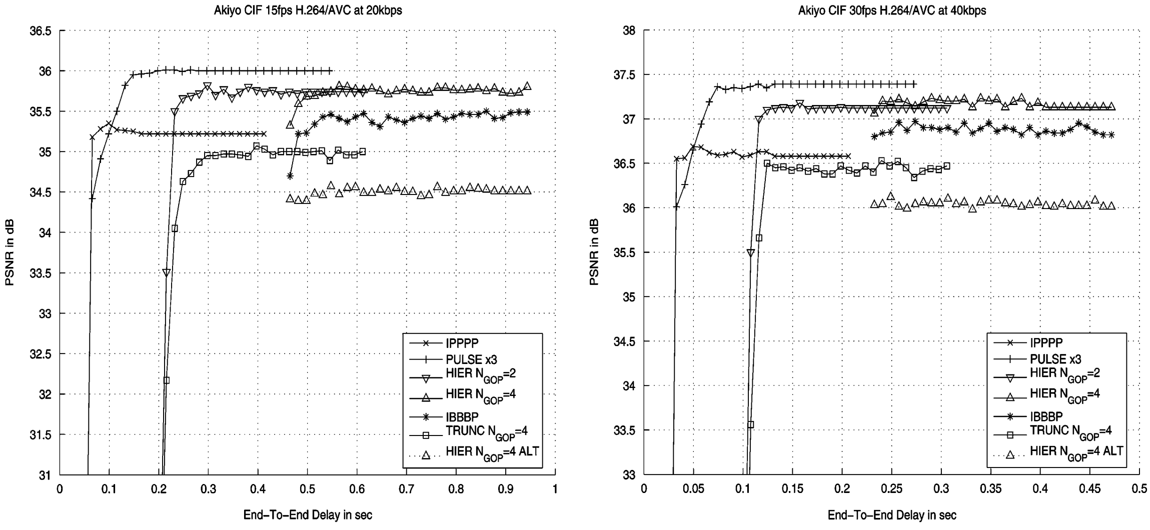
We, therefore, encode the sequence by allocating $R_0 \times c$ bits per frame to frames of level 0, $R_1 \times c$ bits per frame to frames of level 1, and $R_2 \times c$ bits per frame to frames of level 2. Term c is calculated below so that the resulting bit allocation leads to the desired average target rate R_f. Hence, given an average bandwidth constraint of R_f bits per frame we calculate the parameter c as follows:

$$c = \frac{2^L R_f}{R_0 + \sum_{i=1}^{L} 2^{i-1} R_i}. \tag{19}$$

This model is also used to derive the rate allocations for TRUNC and IBBBP.

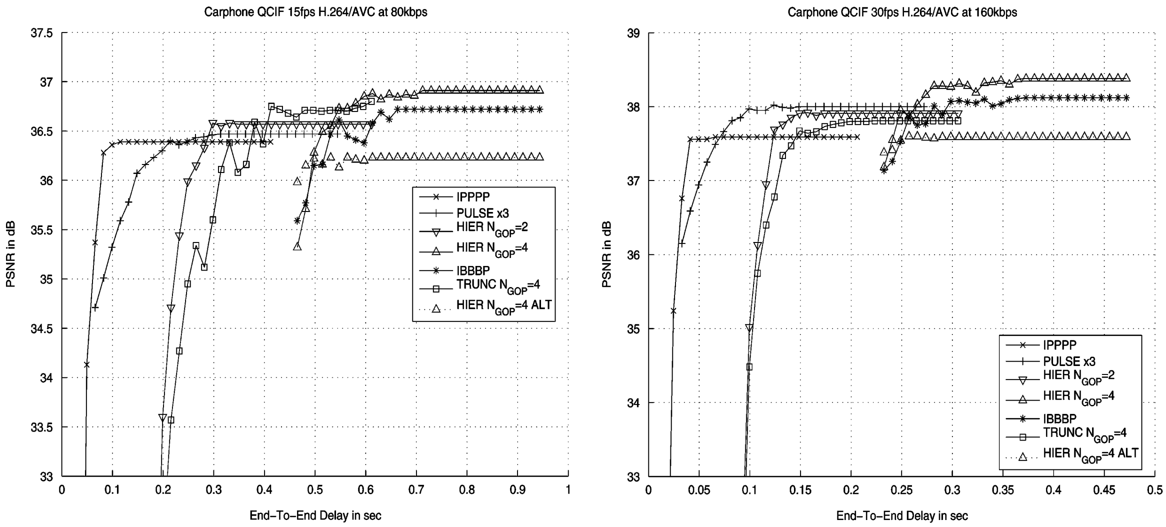
For the PULSE codec the optimal rate allocation varies according to the image sequence. Through limited experimentation, we came up with an empirical rate allocation. We allocate three times the rate for the long-term periodic frames compared to the rate allocated to the regular short-term frames and set the updating parameter $U = 5$. An optimal rate allocation, that adjusts U and the pulsing ratio, would involve buffering a number of frames equal to the updating period and analyzing their correlations, but this is very complex for our intended applications. The factor of three was found as a good all around value through experiments with widely used test sequences. The exact number of bits is calculated adaptively so that the overall rate constraint is satisfied. Better performance could be achieved by optimizing

[2] In reality, the rate allocation for each higher level (with smaller index) has a direct effect on the R-D curve of the immediately lower level as shown in [25]. The R-D curve is shifted for different quantization levels (or bit rate allocations) of its reference frames.



(a) (b)

Fig. 7. PSNR versus delay (in seconds) for fixed source coding bit rate. (a) “Akiyo” CIF 352 × 288 at 15 fps. Initial QP 39. (b) “Akiyo” CIF at 30 fps. Initial QP 30.



(a) (b)

Fig. 8. PSNR versus delay (in seconds) for fixed source coding bit rate. (a) “Carphone” QCIF 176 × 144 at 15 fps. Initial QP 31. (b) “Carphone” QCIF at 30 fps. Initial QP 29.

these parameters, but exploring this large parameter space is beyond the scope of this paper.

An alternative approach to calculate the rate ratios was proposed by a reviewer for the HIER structure with $N_{\rm GOP} = 4$. Assuming that R_t is the target bit rate we write down the following expression:

$$4 \times R_t = R_0 + R_1 + 2 \times R_2. \tag{20}$$

Using (18), this can be rewritten to

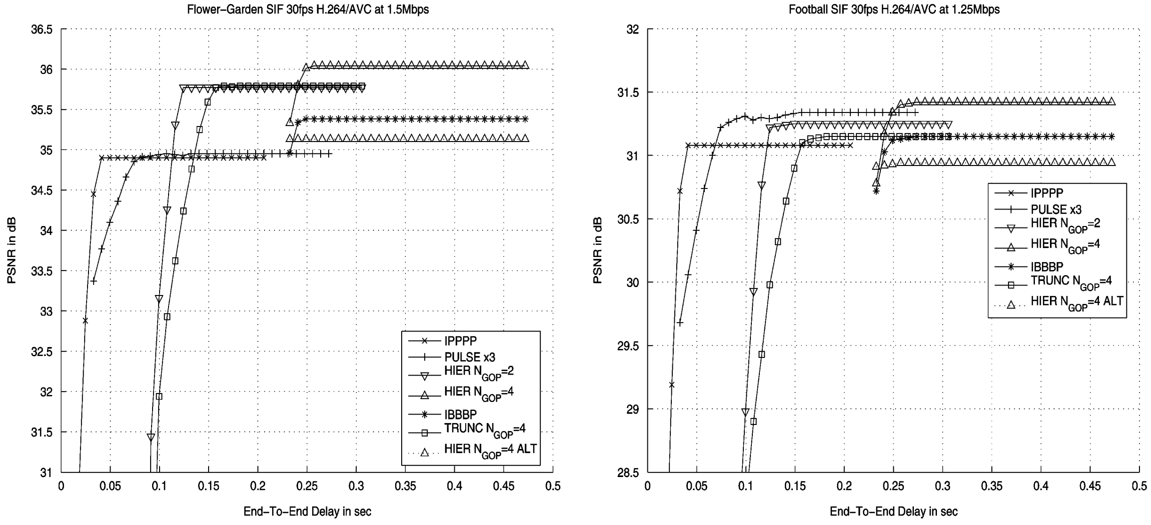
$$4\frac{1}{2}\log_2\frac{\sigma_e^2(1,1)}{D_t(1+\epsilon)} = \frac{1}{2}\log_2\frac{\sigma_e^2(4,1)}{D_r(1+\epsilon)} + \frac{1}{2}\log_2\frac{\sigma_e^2(2,2)}{D_r(1+2\epsilon)} + 2\frac{1}{2}\log_2\frac{\sigma_e^2(1,2)}{D_r(1+3\epsilon)}. \tag{21}$$

By constraining the target rate to a value R_t that corresponds to a distortion D_t, we can then solve the above equation and find D_r. After D_r has been calculated, the calculation of the R_0, R_1, and R_2 is straightforward. We used the ratios obtained this way to re-encode the sequence using the HIER structure and $N_{\rm GOP} =$ 4 and the results were very similar. The primary reason is that the σ_e^2 is not individually optimized for each sequence (and even bit rate). However, accurate estimation of the σ_e would benefit the above approach.

VII. Rate Control

For the IPPPP codec, the rate control algorithm included in the JM 10.1 reference software (described in detail in [15]) is directly used.

For the PULSE codec, the rate control is similar to that in [15] with some critical modifications. We do not allocate rate



(a) (b)

Fig. 9. PSNR versus delay (in seconds) for fixed source coding bit rate. (a) "Flower-Garden" SIF 352 × 240 at 30 fps. Initial QP 29. (b) "Football" SIF at 30 fps. Initial QP 33.

to ST and LT frames from a common budget. The budget is divided into two bins: the ST and the LT rate bins. Two separate rate control "paths" for ST and LT frames draw bits from the respective bins. Each rate control path updates and stores its *own* quadratic rate control model. Using the same quadratic model to control the rate of ST and LT frames is highly ineffective due to coding statistics contamination. Both rate control paths however share the constraint on the encoder buffer status taking care to avoid a buffer overflow.

In each rate control path, the buffer limit is enforced both a) by modifying the QP so as to achieve the target rate but also b) by the last-resort measure of forcing SKIP coding modes on blocks when the buffer is about to overflow. We switch the QP on a basis (*basic unit*) of 11 MBs. The quadratic model of [15] selects a QP for this *basic unit* that *ought* to avoid an overflow. Since the quadratic model is only an estimate, there are cases where SKIP modes must be invoked to avoid overflows. To predict these cases and act accordingly, we assume that signalling a SKIP mode costs two bits, even though in reality the use of CAVLC reduces the actual cost further. In the case of a SKIP mode, the reconstructed MB is a motion-compensated prediction from a previous reference frame. The motion vector is obtained through spatial prediction of neighboring motion vectors.

Concerning the HIER coder, we note that the JM 10.1 reference software includes a scheme [15] that does not explicitly control the rate for B-coded pictures. It was not designed with hierarchical motion-compensation structures in mind. While the rate of the P-coded pictures is strictly controlled by changing the QP in terms of basic units, the B-coded pictures are allocated a single QP value for the entire frame. Thus, the rate is not explicitly controlled. In general, for constant QP allocation, a B-coded picture will be noticeably smaller than its neighboring P-coded pictures, due to the efficiency of bidirectional prediction. Furthermore, the rate control of [15] allocates a QP incremented by two over the average QP of the neighboring P-coded pictures. The B-coded picture is usually smaller than the neighboring P-coded pictures. However, allocating a single QP for the entire frame is a delay nightmare: there is no guarantee on how large or small the B-coded picture will be. As a result, and in a departure from previous literature, we apply explicit rate control to B-coded pictures to ensure that our delay constraint is met.

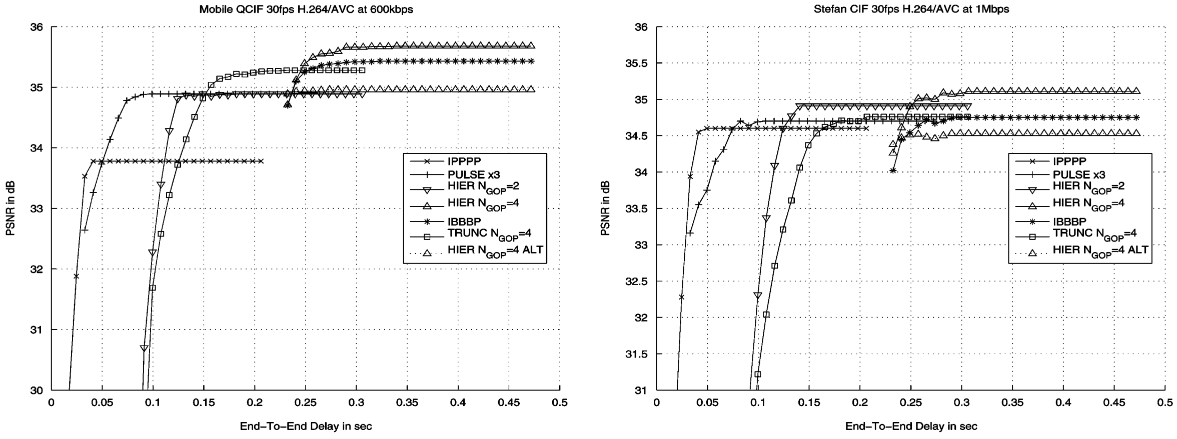
To ensure accurate rate control under tight delay constraints, we adopt the rate control approach of the PULSE codec, with multiple rate-control paths, each of which maintains its own quadratic model. For a hierarchical stream, the number of rate control bins is equal to the number of temporal decomposition levels. For example, for $N_{\mathrm{GOP}} = 4$, we obtain three bins: one for the P-coded pictures, a second one for the RB-coded pictures, and a third for the B-coded pictures. Frames draw their bits only from their corresponding rate control bin. Still, as in the PULSE case, all three bins share the same buffer and thus the same constraint. In cases where the rate control is close to triggering a buffer overflow, we strongly increase the QP and force SKIP modes.

VIII. Results

A. Experimental Setup

A single Intra picture was used at the beginning of each sequence for all tested codecs. The drawback is that random access is quite limited; it is possible only at granularities of, say, 300 frames in our simulations. In practical systems it is highly desirable, and we believe that the performance of our rate allocation model will not be affected by the frequent and periodic insertion of intracoded frames. In that case, the rate allocated to Intra pictures could be some predefined value; it is not calculated by our model. An alternative technique to enable random access would be to progressively refresh (mainly) P-coded frames with intracoded macroblocks.

The encoder of the JM 10.1 reference software of the H.264/AVC coding standard was used for our simulations.



(a)

(b)

Fig. 10. PSNR versus delay (in seconds) for fixed source coding bit rate. (a) "Mobile-Calendar" QCIF 176 × 144 at 30 fps. Initial QP 29. (b) "Stefan" CIF 352 × 288 at 30 fps. Initial QP 31.

Weighted Prediction was not used. Loop filtering was enabled for all frames. The entropy coder used was the context-adaptive binary arithmetic coder (CABAC). Furthermore, memory management control operations (picture order count memory management) and reference picture list reordering (reference reordering) commands were used to ensure that the correct frames are employed for motion compensation in the case of hierarchical structures. The number of reference frames was set to two for IPPPP, PULSE and IBPBP, and was four for HIER with GOP size 4, TRUNC, and IBBBP.

For the IPPPP codec, a single short-term reference frame was used for motion compensation. For the PULSE codec, the updating period was chosen as $U = 5$ and two reference frames were used. For all other variants that used B-coded pictures (HIER, TRUNC, and IBBBP), the P-coded pictures used one reference, and the B-coded pictures used one reference frame from each prediction list (past and future). We note that large values of N_{GOP} such as 8 and 16 are possible (the H.264/AVC specification allows up to 16), but preliminary trials showed that the gain in PSNR is small compared to the dramatic increase in end-to-end delay. Still they offer better temporal scalability, error resilience, and bit stream adaptation.

We note that for IPPPP and PULSE, the output frame order is $[0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8]$. For the HIER $N_{\text{GOP}} = 2$ case, the order is [0 2 1 4 3 6 5 8 7]. For HIER $N_{\text{GOP}} = 4$, the order is [0 4 2 1 3 8 6 5 7]. For the truncated GOP size 4 case (TRUNC), the order is identical to that for $N_{\text{GOP}} = 2$. Last, for the IBBBP case, the order is [0 4 1 2 3 8 5 6 7]. All investigated video codecs produce H.264/AVC [21], [22] compliant bit streams. All results in Section VIII were obtained with the JM 10.1 reference decoder to ensure that they can be decoded correctly.

B. Proposed Rate Allocation

The efficiency of the scheme proposed in Section V-C is illustrated in Fig. 15(b) where we present encoding results for three different rate allocations: a) a trivial uniform rate allocation where each frame, irrespective of its temporal level and prediction distance, receives the same number of bits, b) an intuitive allocation where all B-coded pictures, irrespective of temporal level, receive half the rate of the P-coded pictures, and c) our proposed scheme. We observe that the rate allocation model we obtained outperforms the other two heuristic schemes. We note that allocation b) becomes less efficient than allocation a) as the bit rate increases. The performance improvement for allocation c) over the others tends to increase with increasing rate, which is attributed to our high rate assumptions used for deriving the scheme in Section V-C. The advantage of our method is also visible in Figs. 7–10 that investigate the delay tradeoff where we have plotted the performance of heuristic rate allocation denoted by ALT.

C. Delay Tradeoffs

We investigated the performance of the five types of codecs for a variety of video sequences: *Akiyo* is a very static image sequence. *Carphone* includes localized motion of various kinds. Still, the majority of the activity is due to the instability of the camera inside the car. There is repetitive translational global motion. *Flower* also has high frequency content, and the motion is global and follows mainly the affine model (more complex than translational). *Football* is extremely active with local object motion. *Mobile* has substantial high frequency content and the motion is mostly global due to the horizontal camera pan (translational motion). *Stefan* is a sports clip featuring a tennis court with very high motion.

In Figs. 7–10, we show video quality versus end-to-end delay. The bit rate is fixed for all curves displayed within a graph of those figures. The delay was varied by allocating different numbers of bits to the encoder output buffer (B). Performance increases with delay and GOP size. $N_{\text{GOP}} = 4$ outperforms $N_{\text{GOP}} = 2$, which in turn outperforms $N_{\text{GOP}} = 1$, both IPPPP and PULSE, for *most cases*. The truncated GOP size 4 mostly underperforms $N_{\text{GOP}} = 2$ as expected. Last, PULSE is better than IPPPP.

The performance for "Akiyo" and "Carphone" at both 15 and 30 fps can be seen in Figs. 7 and 8. We subsampled the original 30 fps sequences and encoded them with the proposed codecs. As can be seen we enforced the same number of bits per frame. As expected, the quality for 15 fps was lower than that for 30 fps

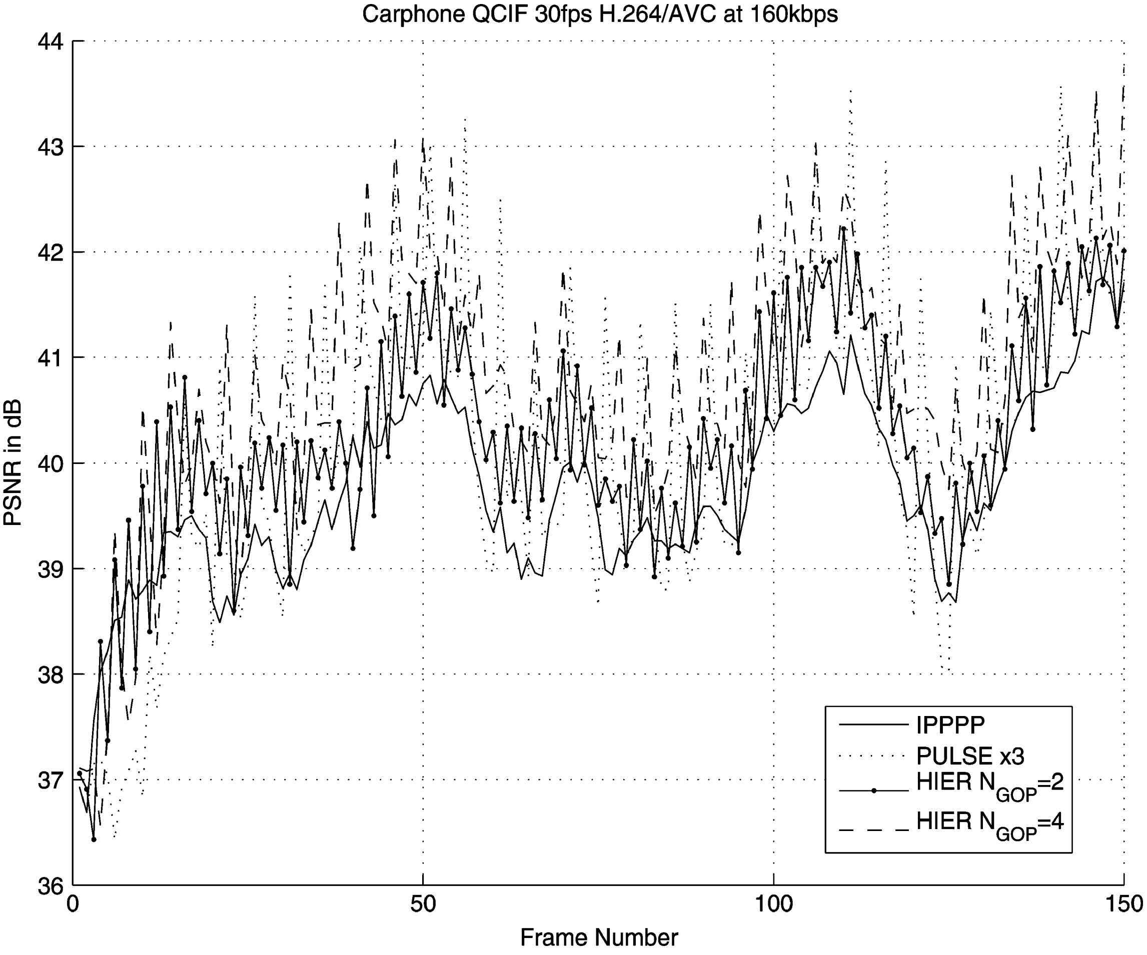


Fig. 11. PSNR versus frame number for various encoders. “Carphone” QCIF 176 × 144 at 30 fps.

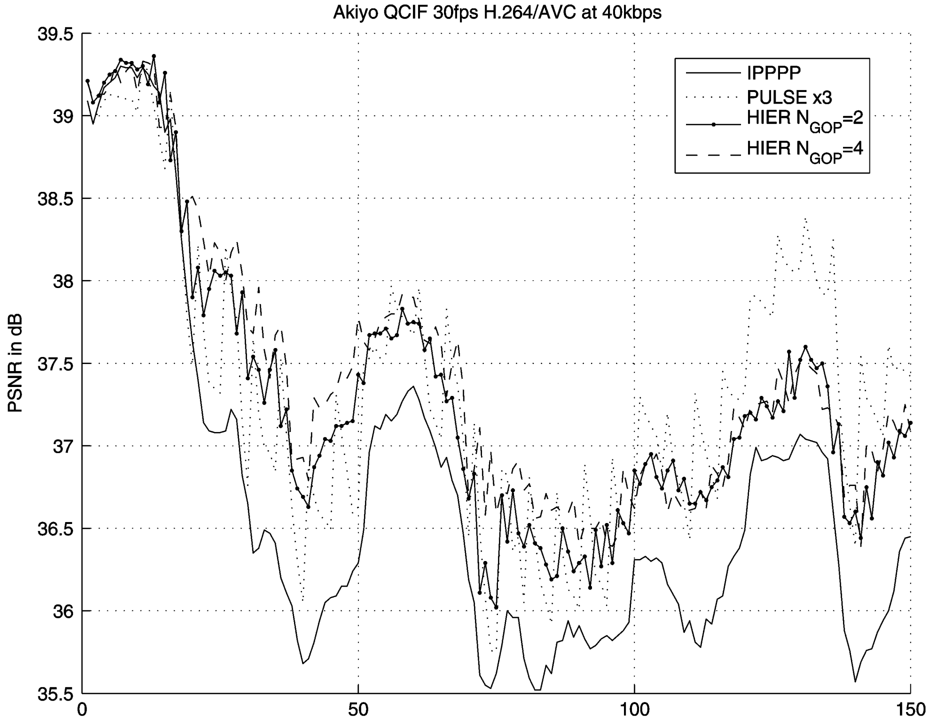
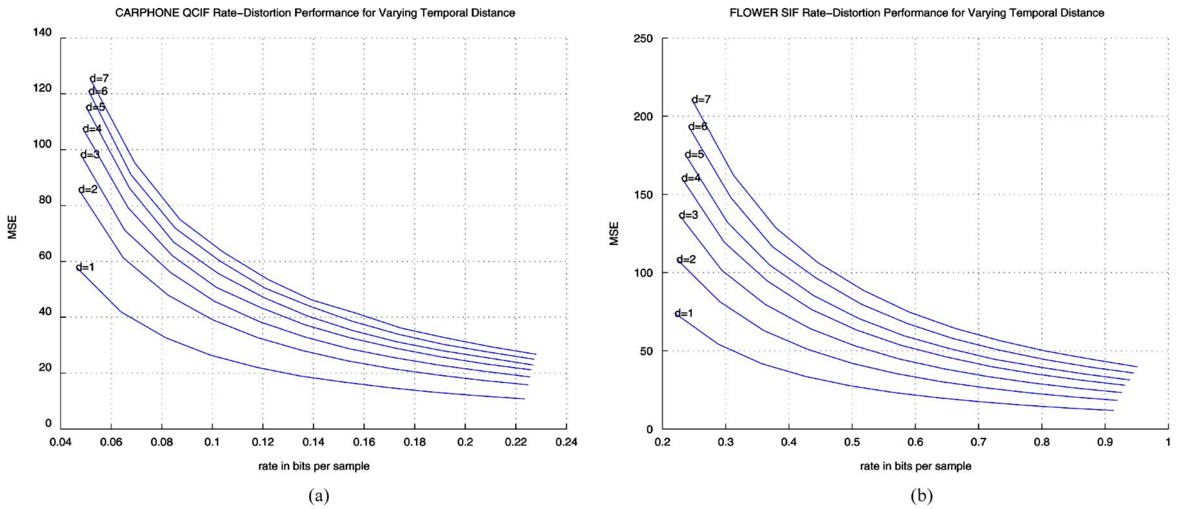


Fig. 12. PSNR versus frame number for various encoders. “Akiyo” CIF 352 × 288 at 30 fps.



(a) (b)

Fig. 13. PSNR versus rate (in bits per sample) for varying temporal prediction distance. (a) “Carphone” QCIF 176 × 144. (b) “Flower” SIF 352 × 240.

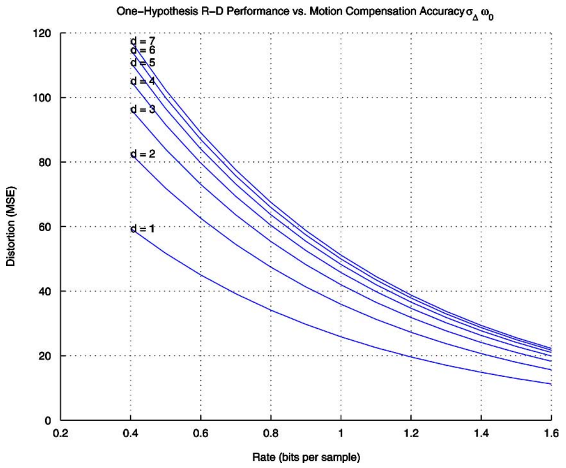


Fig. 14. Theoretical performance versus bit rate for varying $\sigma_\Delta \omega_0$.

since the increased temporal prediction distance decreases the motion-compensation accuracy. Our model performed well in both frame rate cases.

In Fig. 7, we notice that while going from $N_{\rm GOP} = 2$ to $N_{\rm GOP} = 4$ helps performance, the biggest gains come when PULSE is used. Indeed, Akiyo is the only sequence where PULSE outperforms all other codecs. The reason is the static nature of this sequence. For the “Carphone” sequence in Fig. 8, the result is more typical, where the HIER codecs outperform the PULSE codec. The same can be said for “Flower” and “Mobile” in Figs. 9(a) and 10(a) where the performance delta for large GOP is even larger. For the sequences “Football” and “Stefan” in Figs. 9(b) and 10(b) the coding gains are small due to the high motion that is inherent in these sequences.

The numbers we give below for delays are for the 30 fps sequences. We observe in Figs. 7–10 that the IPPPP codec achieves good performance at a delay of around 40 ms. We note that the minimum delay in our system is the transmission time $t_{\rm fr} = 33$ms. The PULSE codec achieves good performance at a delay of around 80 ms. We note that the minimum delay that guarantees good performance can be calculated from U and the long-term to short-term bit budget ratio. The PSNR performance depends on the image sequence. For the highly active “Flower,” “Football,” and “Stefan” there is no gain over the IPPPP codec. Significant gains are however observed for the “Akiyo,” “Carphone,” and “Mobile” sequences.

Moving to the $\rm GOP = 2$ case, we observe that the end-to-end delay needs to be at least 80 ms for good performance. There is no performance gain over PULSE for “Carphone.” However, impressive gains are observed in “Flower.” It is thus evident that hierarchical structures can be very advantageous in static sequences or sequences with global motion.

The TRUNC $\rm GOP = 4$ codec exhibits considerably lower end-to-end delay compared to the HIER $\rm GOP = 4$ codec due to the removal of the backward prediction branch, but the delay is still somewhat higher than the $\rm GOP = 2$ codec. Even though the structural delay is indeed the same, the truncated $\rm GOP = 4$ codec suffers from the fact that the anchor P-coded pictures (e.g., $0, 4, 8, \ldots$) have to be afforded more rate compared to the inner P-coded pictures (e.g., $2, 6, 10, \ldots$) because they are predicted from a larger temporal distance. The rate allocation scheme we developed in Section V-C gives us an approximate estimate of the required increase in rate to compensate for the drop in motion compensation efficiency. From the graphs we see that this structure is not beneficial in terms of rate versus distortion performance. However, it provides additional temporal scalability for much lower delay compared to $\rm GOP = 4$. Regarding the IBBBP prediction structure we observe that in terms of delay it is closer to that of $\rm GOP = 2$, which is a very desirable property. Recall that the decoding delay of such a structure does not increase with $N_{\rm GOP}$, in contrast to the HIER codecs. However, the performance is not good, and apart from potential problems with our model when used for these structures, another reason for this finding might be the fact that we do not use weighted prediction.

The increase of the GOP size to 4 increases delay considerably to almost 270 ms. Apart from increased GOP delay, the anchor P-coded pictures get large, contributing further to delay. The three B-coded pictures in each GOP need many fewer bits to be encoded. In terms of performance gain, “Carphone” and

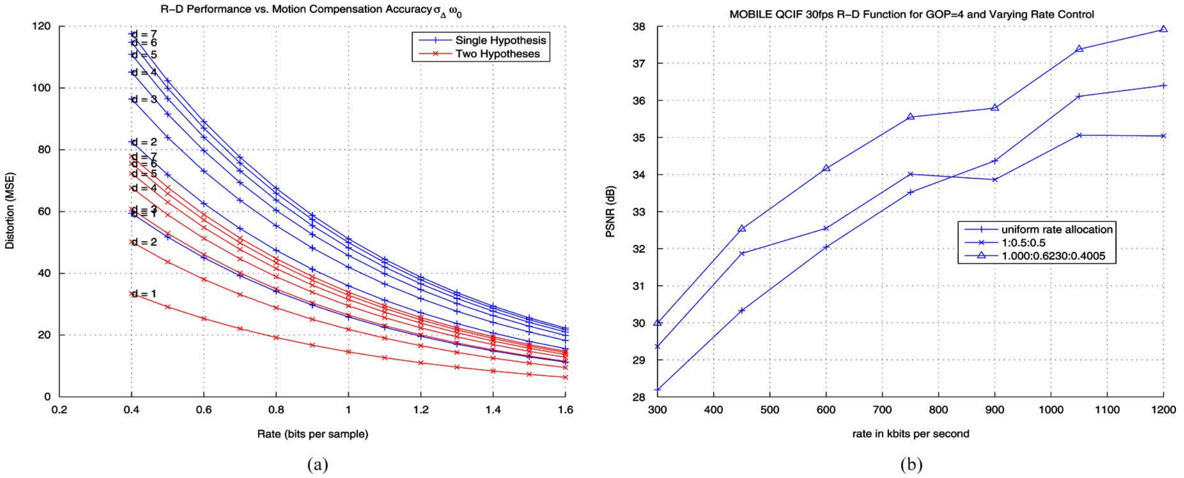


Fig. 15. (a) Single and double hypothesis R-D behavior calculated theoretically for difference temporal prediction distances. (b) Performance of the proposed rate allocation for "Mobile" QCIF at 30 fps with $N_{\text{GOP}} = 4$.

"Mobile" benefit the most. The large gain in "Mobile" is attributed not only to its global motion but also to the fact that it is translational.

Last, in Figs. 11 and 12, we plot the performance of four codecs for each frame in the sequence. We plot only 150 frames for ease of presentation. For the "Akiyo" sequence we can see some very limited pulsing of the PSNR values and it is obvious why the HIER and PULSE perform so well. The curve for IPPPP is almost always below the lowest point of the other curves. For "Carphone" there are some variations of 1–2 dB in PSNR but again the lowest points in the curves of the HIER codecs are above the IPPPP curve. For PULSE the variation is higher and the pulsing ratio might have to be adjusted.

IX. Conclusion

We studied end-to-end delay versus compression efficiency tradeoffs for video encoders with varying GOP size. The end-to-end delay depends on the encoder input and output buffer delay. The output buffering delay was found to be a function of the rate allocation. We hence investigated the effect on delay of allocating more bits to some frames than the rest. All these codecs are H.264/AVC compliant. We implemented a robust rate control algorithm for the PULSE codec as well as for the hierarchical B-pictures. The work in [7] used constant QPs without any consideration of rate or delay constraints. Here, in a departure from previous work, we operated under both constraints.

A theoretical framework was derived for rate allocation in the context of varying temporal distances and number of prediction hypotheses. We found that the standard deviation of the motion compensation displacement error σ_Δ varies approximately *linearly* with the temporal prediction distance Δ_t.

We investigated constraints in structural delay through prediction branch truncation for lower delay. This leads to worse compression efficiency but is efficient in terms of scalability and bit stream adaptability. Our rate allocation scheme was used to find an efficient bit distribution for these cases.

The study of the delay tradeoffs yielded the following conclusions.

a) IPPPP performs well for low delay applications and for sequences with high motion.
b) PULSE is advantageous for relatively static sequences with repetitive content.
c) $N_{\text{GOP}} > 1$ structures benefit from static sequences and from sequences with global motion.
d) As N_{GOP} increases, the gain is nontrivial only if the sequence is either static, or if the global motion is translational.
e) For the sequences we evaluated, the delay thresholds are as follows: between 40 and 80 ms, IPPPP is the best choice, between 80 and 125 ms PULSE performs well, the large space between 125 and 270 ms is dominated by $N_{\text{GOP}} = 2$, and for delays larger than 270 ms, then $N_{\text{GOP}} = 4$ is the best choice. Delays larger than 270 ms are only however useful in cases of live event broadcast or streaming of stored content. They are prohibitive for real-time interactive communication.
f) The truncated GOP $= 4$ codec underperforms the GOP $= 2$ codec but has similar delay with the added advantage of increased temporal scalability.

References

[1] A. Leontaris and P. C. Cosman, "End-to-end delay for hierarchical B-pictures and pulsed quality dual frame video coders," in *Proc. IEEE Int. Conf. Image Processing*, Oct. 2006, pp. 3133–3136.

[2] G. Pau, B. Pesquet-Popescu, M. van der Schaar, and J. Vieron, "Delay-performance trade-offs in motion-compensated scalable subband video compression," presented at the Advanced Concepts for Intelligent Vision Systems, Sep. 2004.

[3] J. Reichel, H. Schwarz, and M. Wien, "Scalable video coding working draft 1," presented at the Joint Video Team JVT of ISO/IEC MPEG and ITU-T VCEG, JVT-N020, Jan. 2005.

[4] G. Pau, J. Vieron, and B. Pesquet-Popescu, "Video coding with flexible MCTF structures for low end-to-end delay," in *Proc. IEEE Int. Conf. Image Processing*, Sep. 2005, vol. 3, pp. 241–244.

[5] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MCTF," in *Proc. IEEE Int. Conf. Multimedia and Expo*, Jul. 2006, pp. 1929–1932.

[6] H. Shen, X. Sun, F. Wu, and S. Li, "Rate-distortion optimization for fast hierarchical B-picture transcoding," in *Proc. IEEE Int. Symp. Circuits and Systems*, May 2006, pp. 5279–5282.

[7] A. Leontaris, V. Chellappa, and P. C. Cosman, "Optimal mode selection for a pulsed-quality dual frame video coder," *IEEE Signal Process. Lett.*, vol. 11, no. 12, pp. 952–955, Dec. 2004.

[8] H. Schwarz, D. Marpe, and T. Wiegand, "Comparison of MCTF and closed-loop hierarchical B pictures," Joint Video Team of ISO/IEC MPEG and ITU-T VCEG, JVT-P059, Jul. 2005.

[9] *Test model 5*, JTC1/SC29/WG11 Coding of Moving Pictures and Associated Audio, ISO-IEC/JTC1/SC29/WG11, 1994.

[10] P. Westerink, R. Rajagopalan, and C. Gonzales, "Single-pass constant- and variable-bit-rate MPEG-2 video compression," *IBM J. Res. Develop.*, vol. 43, no. 4, Jul. 1999.

[11] S. Ma, W. Gao, and Y. Lu, "Rate control on JVT standard," Joint Video Team of ISO/IEC MPEG and ITU-T VCEG, JVT-D030, Jul. 2002.

[12] S. Ma, W. Gao, Y. Lu, and D. Zhao, "Improved rate control algorithm," Joint Video Team of ISO/IEC MPEG and ITU-T VCEG, JVT-E069, Oct. 2002.

[13] T. Chiang and Y. Zhang, "A new rate control scheme using quadratic rate distortion model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 4, pp. 287–311, Apr. 1997.

[14] Z. G. Li, F. Pan, K. P. Lim, G. Feng, X. Lin, and S. Rahardja, "Adaptive basic unit layer rate control for JVT," Joint Video Team of ISO/IEC MPEG and ITU-T VCEG, JVT-G012, Mar. 2003.

[15] K.-P. Lim, G. Sullivan, and T. Wiegand, "Text description of joint model reference encoding methods and decoding concealment methods," Joint Video Team of ISO/IEC MPEG and ITU-T VCEG, JVT-K049, Mar. 2004.

[16] B. Girod, "The efficiency of motion-compensating prediction for hybrid coding of video sequences," *IEEE J. Sel. Areas Commun.*, vol. SAC-5, no. 8, pp. 1140–1154, Aug. 1987.

[17] ——, "Motion-compensating prediction with fractional-pel accuracy," *IEEE Trans. Commun.*, vol. 41, no. 4, pp. 604–612, Apr. 1993.

[18] A. R. Reibman and B. G. Haskell, "Constraints on variable bit-rate video for ATM networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 2, no. 4, pp. 361–372, Dec. 1992.

[19] J. Ribas-Corbera, P. A. Chou, and S. L. Regunathan, "A generalized hypothetical reference decoder for H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 674–687, Jul. 2003.

[20] M. Isnardi, MPEG-2 Video Compression, SMPTE Tutorial Overview, Sarnoff Corporation, Nov. 1999.

[21] *Advanced Video Coding for Generic Audiovisual Services*, [Online]. Available: http://www.itu.int/rec/T-REC-H.264/, ITU-T Recommendation H.264

[22] G. Sullivan, T. Wiegand, D. Marpe, and A. Luthra, Text of ISO/IEC 14496-10 Advanced Video Coding 3rd edition, ISO/IEC JTC 1/SC 29/WG11 N6540, Jul. 2004.

[23] M. Karczewicz and Y. Bao, "Need for further AVC Test model enhancements," Joint Video Team of ISO/IEC MPEG and ITU-T VCEG, JVT-L034, Jul. 2004.

[24] B. Girod, "Efficiency analysis of multihypothesis motion-compensated prediction for video coding," *IEEE Trans. Image Process.*, vol. 9, no. 2, pp. 173–183, Feb. 2000.

[25] K. Ramchandran, A. Ortega, and M. Vetterli, "Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders," *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 533–545, Sep. 1994.

[26] T. Rusert, K. Hanke, and J.-R. Ohm, "Transition filtering and optimized quantization in interframe wavelet video coding," in *Proc. SPIE Visual Communications and Image Processing*, Jul. 2003, pp. 682–693.

[27] A. Mavlankar and E. Steinbach, "Distortion prediction for motion-compensated lifted Haar wavelet transform and its application to rate allocation," presented at the Int. Picture Coding Symposium PCS, Dec. 2004.

[28] M. Wang and M. van der Schaar, "Rate-distortion modeling for wavelet video coders," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Mar. 2005, vol. 2, pp. 53–56.

[29] T. Rusert and J.-R. Ohm, "Macroblock based bit allocation for SNR scalable video coding with hierarchical B pictures," presented at the IEEE Int. Conf. Image Processing, Oct. 2006.

[30] N. Farvardin and J. W. Modestino, "Rate-distortion performance of DPCM schemes for autoregressive sources," *IEEE Trans. Image Process.*, vol. 31, no. 3, pp. 402–418, May 1985.

[31] P. Ramanathan and B. Girod, "Rate-distortion analysis for light field coding and streaming," presented at the EURASIP Signal Processing: Image Communication, Nov. 2005.

[32] U. Horn, T. Wiegand, and B. Girod, "Bit allocation methods for closed-loop coding of oversampled pyramid decompositions," in *Proc. IEEE Int. Conf. Image Processing*, Oct. 1997, vol. 2, pp. 17–20.

[33] M. Flierl and B. Girod, "Video coding with motion compensation for groups of pictures," in *Proc. IEEE Int. Conf. Image Processing*, Sep. 2002, vol. 1, pp. 69–72.

Athanasios Leontaris (S'97–M'06) received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2000, and the M.S. and Ph.D. degrees in electrical engineering from the University of California at San Diego (UCSD), La Jolla, in 2002 and 2006, respectively.

From 2000 to 2001, he was a Research Associate with the Informatics and Telematics Institute, Greece. He was a summer intern at AT&T Labs—Research, New Jersey and at NTT Network Innovation Labs, Japan in 2004 and 2005, respectively. Currently, he is a Senior Research Engineer at Dolby Laboratories, Inc., Burbank, CA. His research interests include image and video compression, video communication, multimedia processing, and image quality modeling.

Pamela C. Cosman (S'88–M'93–SM'00) received the B.S. degree (with honors) in electrical engineering from the California Institute of Technology, Pasadena, in 1987, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1989 and 1993, respectively.

She was an NSF postdoctoral fellow at Stanford University and a Visiting Professor at the University of Minnesota during 1993–1995. In 1995, she joined the faculty of the Department of Electrical and Computer Engineering, University of California, San Diego, where she is currently a Professor and Director of the Center for Wireless Communications. Her research interests are in the areas of image and video compression and processing.

Dr. Cosman is the recipient of the ECE Departmental Graduate Teaching Award (1996), a Career Award from the National Science Foundation (1996–1999), and a Powell Faculty Fellowship (1997–1998). She was a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS June 2000 Special Issue on Error-Resilient Image and Video Eoding, and was the Technical Program Chair of the 1998 Information Theory Workshop, San Diego. She was an Associate Editor of the IEEE COMMUNICATIONS LETTERS (1998–2001), and an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS (2001–2005). She was a Senior Editor (2003–2005), and is now the Editor-in-Chief, of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. She is a member of Tau Beta Pi and Sigma Xi.