Title
Optimal Mode Selection for a Pulsed-Quality Dual-Frame Video Coder

Permalink
https://escholarship.org/uc/item/2338k8hv

Journal
IEEE Signal Processing Letters, 11(12)

ISSN
1070-9908

Authors
Leontaris, A.
Chellappa, V.
Cosman, P. C.

Publication Date
2004-12-01

DOI
10.1109/LSP.2004.838197

Peer reviewed

Optimal Mode Selection for a Pulsed-Quality Dual-Frame Video Coder

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

***Abstract*—A dual-frame video coder employs two past reference frames for motion compensated prediction. Compared to conventional single frame prediction, the dual-frame encoder can have advantages both in distortion-rate performance and in error resilience. In previous work, it was shown that optimal mode selection can enhance the performance of a dual-frame encoder. In another strand of previous work, it was shown that uneven assignment of quality to frames, to create high-quality (HQ) long-term reference frames, can enhance the performance of a dual-frame encoder. In this letter, we combine these two strands and demonstrate the performance advantages of optimal mode selection among HQ frames for video transmission over noisy channels.**

***Index Terms*—Dual-frame buffer, high-quality updating, H.264, mode switching, multiple frame prediction, per-pixel estimation, video compression.**

I. INTRODUCTION

TRADITIONALLY, hybrid video codecs employ motion-compensated prediction to compress an input raw video stream. A block of pixels in the current frame is predicted from a displaced block in a previous frame. A motion vector points to the coordinates of the displaced block. The difference (error signal) between the original block and its prediction is compressed and transmitted along with the corresponding displacement (motion) vectors. This approach has formed the cornerstone of modern video coding algorithms such as MPEG-4 and H.263+.

Performance was improved when the search for the best prediction included additional past frames apart from the previous one. Examples of this multiple reference frame approach are [1]–[3], and it has recently been standardized as part of the H.264/AVC video coding standard [4]. To counter the increased memory and computational cost, the number of reference frames can be constrained to be small. Fukuhara *et al.* used only two reference frames [5]. The first reference buffer contained the previous frame, and the second one contained a reference frame from the distant past that was periodically updated. We refer to this as *dual-frame* coding.

A novel algorithm for estimating distortion due to packet losses was introduced in [6], for conventional single-frame encoding. This distortion estimation was extended in [7], to allow dual-frame coding with rate-distortion optimal mode switching. For packet loss scenarios, the performance increased noticeably. In other recent work [8], periodic high-quality (HQ) frames were generated, and retained as the long-term frames in a dual-frame encoding approach.

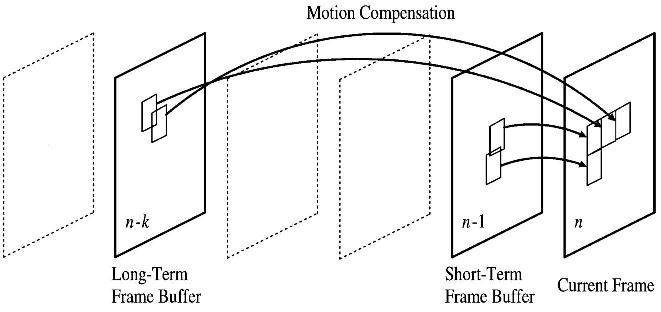

Fig. 1. Dual-frame buffer scheme.

In this letter, we combine the HQ construction and buffering of long-term frames from [8] with the optimal mode selection approach from [7] and show that the combination provides a significant advantage for lossy packet networks. For reasons of simplicity and reduced computational complexity, we made use of half-pel motion compensation to allow easier calculation of the distortion estimate, and the loop filter was disabled as well. This letter is organized as follows. Section II describes how distortion was estimated and optimal mode selection was performed for a dual-frame encoder. Section III discusses the HQ updating approach. In Section IV, we provide experimental results for the combination of these two approaches. The letter is concluded with Section V.

II. OPTIMAL MODE SELECTION FOR A DUAL-FRAME CODER

Dual-frame motion compensation is depicted in Fig. 1, and works as follows. While encoding frame n, the encoder and decoder both maintain two reference frames in memory. The short-term reference is frame $n-1$. The long-term reference can be selected in a number of ways; we used *jump updating*, in which the long-term reference frame varies from as recent as frame $n-2$ to as old as frame $n-N-1$. When encoding frame n, if the long-term reference frame is $n-N-1$, then, when the encoder moves on to encoding frame $n+1$, the short-term reference frame slides forward by one to frame n, and the long-term reference frame jumps forward by N to frame $n-1$. The long-term reference frame then remains static for N frames, and then jumps forward again. We refer to N as the updating parameter. This approach was first adopted in [5].

In dual-frame motion compensation, each macroblock (MB) can be encoded in one of three coding modes: intra coding, inter

---

Manuscript received March 6, 2004; revised May 26, 2004. This work was supported in part by the National Science Foundation, the Office of Naval Research and by the CoRe program of the State of California. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jean-Christophe Pesquet.

The authors are with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093-0407 USA (e-mail: aleontar@code.ucsd.edu; vchellap@code.ucsd.edu; pcosman@code.ucsd.edu).

Digital Object Identifier 10.1109/LSP.2004.838197

coding using the short-term buffer (inter-ST-coding), and inter coding using the long-term buffer (inter-LT-coding). In [7], the choice among these three was made using an extended version of the ROPE algorithm, as described briefly below.

Using the notation from [6], we use $f_n, \hat{f}_n$, and $\tilde{f}_n$ to denote the original frame n, the encoder reconstruction of the compressed frame, and the decoder version (possibly error concealed) of the frame, respectively. We assume that the long-term frame buffer was updated m frames ago. Thus, it contains $\hat{f}_{n-m}$ at the transmitter and $\tilde{f}_{n-m}$ at the receiver. The expected distortion for pixel i in frame n is

$$\begin{aligned} d_n^i &= E\left\{\left(f_n^i - \tilde{f}_n^i\right)^2\right\} \\ &= \left(f_n^i\right)^2 - 2 f_n^i E\left\{\tilde{f}_n^i\right\} + E\left\{\left(\tilde{f}_n^i\right)^2\right\}. \qquad (1) \end{aligned}$$

Calculation of d_n^i requires the first and second moments of the random variable of the estimated image sequence $\tilde{f}_n^i$. We have two separate inter modes, the *inter-ST* and *inter-LT*. Let i denote the pixel in the current frame, k denote the pixel in the previous frame that is associated with pixel i in the current frame using error concealment, and j denote the pixel in the reference frame (either ST or LT) that is the prediction of pixel i in the current frame derived using the motion vector. Let p denote the packet erasure rate (which equals the pixel loss probability for our variable-length packets which contain a single horizontal group of blocks).

The moments for a pixel in an *intra*-coded MB are [6]

$$\begin{aligned} E\left\{\tilde{f}_n^i\right\} &= (1-p)\left(\hat{f}_n^i\right) + p(1-p)E\left\{\tilde{f}_{n-1}^k\right\} \\ &\quad + p^2 E\left\{\tilde{f}_{n-1}^i\right\} \qquad (2) \\ E\left\{\left(\tilde{f}_n^i\right)^2\right\} &= (1-p)\left(\hat{f}_n^i\right)^2 + p(1-p)E\left\{\left(\tilde{f}_{n-1}^k\right)^2\right\} \\ &\quad + p^2 E\left\{\left(\tilde{f}_{n-1}^i\right)^2\right\}. \qquad (3) \end{aligned}$$

The moments of $\tilde{f}_n^i$ for a pixel in an *inter*-coded MB are

$$\begin{aligned} E\left\{\tilde{f}_n^i\right\} &= (1-p)\left(\hat{e}_n^i + E\left\{\tilde{f}_{n-g}^j\right\}\right) \\ &\quad + p(1-p)E\left\{\tilde{f}_{n-1}^k\right\} + p^2 E\left\{\tilde{f}_{n-1}^i\right\} \qquad (4) \\ E\left\{\left(\tilde{f}_n^i\right)^2\right\} &= (1-p)\left(\left(\hat{e}_n^i\right)^2 + 2\hat{e}_n^i E\left\{\tilde{f}_{n-g}^j\right\}\right. \\ &\quad \left. + E\left\{\left(\tilde{f}_{n-g}^j\right)^2\right\}\right) + p(1-p)E\left\{\left(\tilde{f}_{n-1}^k\right)^2\right\} \\ &\quad + p^2 E\left\{\left(\tilde{f}_{n-1}^i\right)^2\right\} \qquad (5) \end{aligned}$$

where these equations are valid for the *inter-LT* mode if $g = m$ and they are valid for the *inter-ST* mode if $g = 1$. Using these equations, the encoder can estimate recursively the per-pixel distortion of the reconstructed video at the decoder. For more details, we refer the reader to [6] and [7].

The only major difference of the mode selection in this work, compared to [7], is that we use H.264 here, and therefore need to change the approximation used for half-pixel motion vectors, to account for the 6-tap interpolation filters used in H.264.

Given the distortion estimate, the encoder switches between intra, inter-ST or inter-LT coding on a MB basis, in an optimal fashion for a given bit rate and packet loss rate. The goal is to minimize the total distortion subject to a bit rate constraint. Individual MB contributions to this cost are additive, thus it can be minimized on a MB basis. Therefore, the encoding mode for each MB is chosen by minimizing

$$\min_{(\text{mode})} J_{\text{MB}} = \min_{(\text{mode})} \left(D_{\text{MB}} + \lambda R_{\text{MB}}\right)$$

where $D_{\text{MB}} = \sum_{i \in MB} d_n^i$ and R_{MB} denote per MB distortion and rate, respectively, and λ is the Lagrange multiplier. The coding *mode* (*intra*, *inter-ST*, and *inter-LT*) is chosen to minimize the Lagrangian cost. Because of the uneven quality levels (discussed in the next section) being assigned in the current paper, contrary to what was done in [7], we do not optimize over the quantization parameter (QP). We instead use the one chosen for that particular frame (or MB), manually or through a rate allocator.

III. HQ UPDATING

A question that arises when designing such a system is the choice of the optimal update parameter N for a given image sequence, frame rate and bit rate. It depends heavily on the sequence's characteristics, such as occlusion effects and scene changes. An optimal solution will require significant computation. In [8], it was proposed not to attempt to *select* an optimal frame to be buffered as long-term, but rather to *construct* a good frame explicitly. In this approach, every N frames, one frame is coded with additional bit rate at the expense of other regular frames. This frame is then buffered and used as the long-term reference frame for the subsequent N frames.

A second issue is how to allocate bit rate to the long-term frame. In this work, as in [8], the rate allocation was heuristic. We used a quantization parameter for the long-term frame that was lower by seven compared to the quantization parameter for the regular frames.

The actual transmission of extra bits for the periodic high-quality frames can be accomplished in two ways. One can incur extra delay for the HQ frames, by using extra transmission time for them, sending more information at the same average bit rate. In another realization that is not prone to delays, the sender could use extra channel bandwidth for a short period of time, to send the extra bits for the HQ frames.

It was found in [8] that using a dual-frame encoder with periodic HQ frames (pulsed quality) that could be used as long-term frames provided about a 0.6 dB advantage over a regular dual-frame encoder where the long-term reference frames have the same quality as other frames.

IV. EXPERIMENTAL RESULTS

The previous work with pulsed-quality dual frames considered only transmission over noiseless channels [8]. In this letter, our goal is to use the pulsed-quality creation of long-term reference frames, but for packet erasure channels, and to use the op-

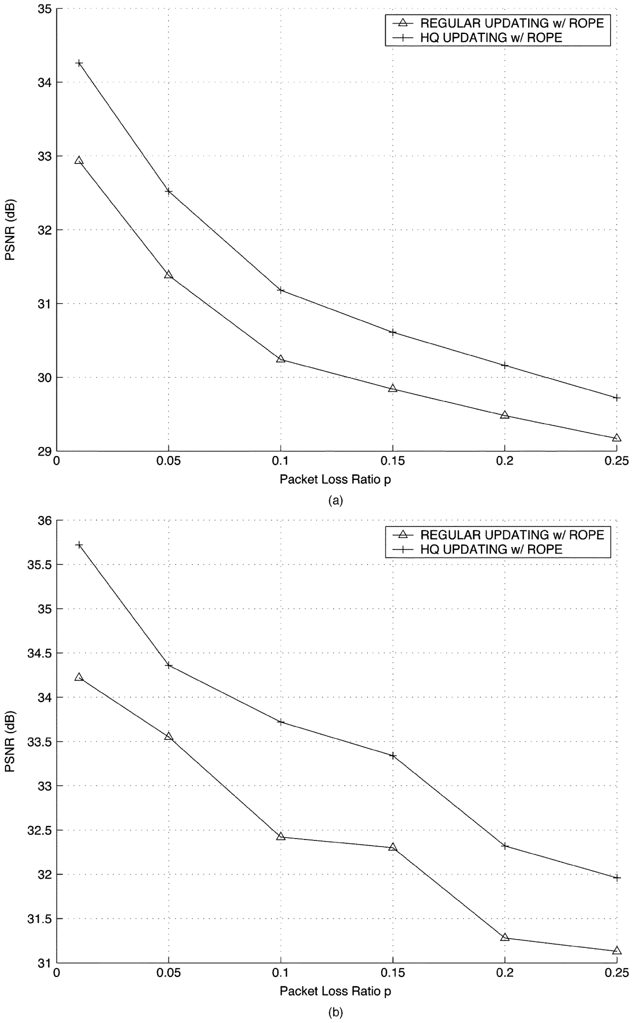

Fig. 2. PSNR performance versus packet loss rate. (a) Image sequence "carphone" QCIF at 10 fps, $N = 10, 122.5$ kbps. (b) Image sequence "mother-daughter" QCIF at 10 fps, $N = 10, 34.4$ kbps.

timal mode selection approach of [7] to optimally select among *intra* coding, *inter-low-quality* coding, and *inter-high-quality* coding based on estimates of the distortion that arises both from packet erasures and from the difference in low and HQ frames.

The dual-frame buffer with HQ updating was implemented with the H.264/AVC reference software version JM 7.4. The encoder was modified, using functionality built into the standard. The resulting video codec produces a fully standard compliant H.264 bitstream. To limit the complexity in implementing the per-pixel estimator from [6], we employed half-pixel motion vectors. For similar reasons, the loop filter was disabled. Since we deal with standard QCIF imagery at 176 × 144 pixels, we took each slice (packet) to contain 11 MBs, so each variable-length packet is equivalent to a horizontal slice of MBs.

A multiple frame buffer of size two was employed. In the case of regular updating, the first frame is the previous and the second a long-term one. When HQ updating is used, the sole difference is that the frames selected to be buffered as long-term ones have been explicitly coded with a lower QP compared to all others. In our simulations we code the long-term frames with a QP that is lower by seven compared to the general QP for the entire sequence. The long-term frame was updated every ten frames (updating parameter $N = 10$). This fixed value was selected experimentally as in [8] and is a compromise between the optimal values for many sequences.

Packet losses corrupt the bitstream. Packets can be decoded independently of one another. Error concealment is applied by using the median of the motion vectors of the three upper MBs to conceal from the previous frame. If the upper slice has been lost as well, then we just copy the co-located MB from the previous frame. The error concealment is modeled within the distortion estimation equations.

One hundred different random patterns were used to obtain the displayed results. Since the regular dual-frame coder takes no account of the possibility of packet losses in choosing between the coding modes, it underutilizes the intra mode, and performs very poorly in a lossy environment. Thus, we also investigated the performance of random intra refresh algorithms.

The QP value was selected empirically in [8] after extensive experimentation for standard video sequences. Constraining the QP selection and comparing against intra-coded long-term frames, we found that the present scheme outperformed intra-LTs for most of the cases by up to 1 dB.

Fig. 3(a) illustrates the system's peformance for the "carphone" image sequence, which consists of a man talking on his videophone in a moving car. For a packet loss rate of 10% the "zero intra" coding proves extremely error-prone. It has very few intra-coded MBs, and a $p = 10\%$ packet loss ratio drops the PSNR down to 21 dB from more than 32 dB. Increasing the allocated rate has no effect since the absence of intra-coded MBs totally compromises the bitstream's resilience.

Attempting to protect the bitstream by employing some intra-coded MBs, we used a random intra refresh update. We experiment by forcing 20, 33, and 45 random intra-coded MBs *per-frame*. Performance increases with the transmission rate, due to the added protection of the intra MBs, particularly as the number of intra-coded MBs increases.

However, we observe that providing HQ to the long-term frames does less well than regular quality for these heuristic intra refresh approaches. This is likely due to the fact that the HQ frames are depriving other frames of their share of rate, and the random intra refresh MBs also deprive other MBs of their share of rate; the competing effects of these heavy rate users hurt the final performance. Similar conclusions can be drawn in Fig. 3(b).

However, with the use of rate-distortion optimal mode selection among *intra* coding, *inter-low-quality* coding, and *inter-high-quality* coding, the pulsed-quality dual-frame approach outperforms regular dual-frame updating by as much as 1–1.5 dB throughout Fig. 3(a) and (b). For Fig. 3(b) in particular, the performance gap increases with the bit rate.

We then investigated the performance of the system for varying packet loss ratios. Fig. 2(a) and (b) demonstrate that HQ updating holds a comfortable lead over regular updating for packet loss ratios ranging from 1% to 25%. The performance gain varies from 0.6–1.5 dB, depending on the image sequence characteristics. It is in general higher for low-motion sequences such as "mother-daughter" but still can reach 1 dB for active sequences such as "carphone". The QPs were chosen so as to achieve the same ($\pm 5\%$) total bit rate for the graph points.

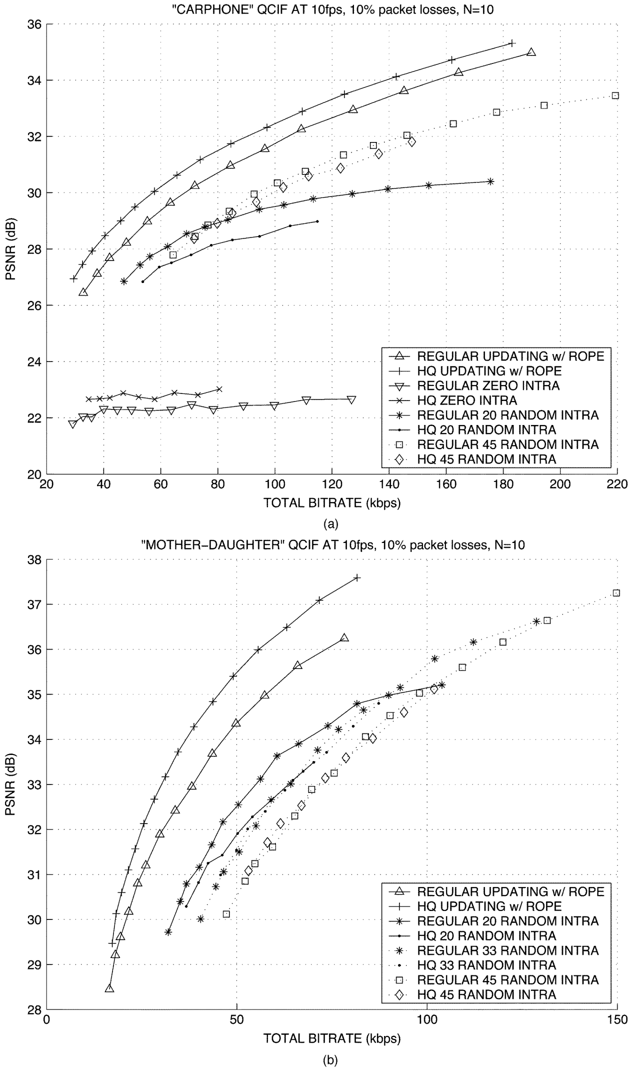

Fig. 3. Packet loss ratio $p = 10\%$. (a) Image Sequence "carphone". (b) Image Sequence "mother-daughter".

We experimented with using more than two reference frames, and found, as in [9], that expanding the reference buffer size beyond two frames produces sharply diminishing returns. In particular, we tried: 1) two long-term (high-quality) frames plus one short-term frame; 2) two short-term frames and one long-term (HQ) frame; and 3) one long-term and four short-term frames. In all cases, the gains over the dual-frame high-quality case were quite small (0.1–0.2 dB). It appears that the immediate past frame captures most of the benefit that short-term references can provide (high correlation with current frame), and a single high-quality frame captures most of the benefit that the high-quality long-term past frame can provide.

Subjective quality evaluation showed that the periodic coding of frames at HQ is only rarely noticeable for the error-free case, and for the error-prone case, it is completely masked by the error concealment and propagation.

V. CONCLUSION

In conclusion, the dual-frame coder with pulses of HQ provided to the long-term frame, when used in conjunction with random intra refresh, performs *less well* than the regular dual-frame coder where long-term frames are chosen from among regular quality frames. The optimal mode selection does significantly better than random intra refresh for both regular quality and pulsed-quality dual-frame coders, and the pulsed-quality approach performs *much better* than the regular dual-frame approach when used with the optimal mode selection. This result says that the method of creating or choosing a long-term reference frame (pulsed quality or regular quality) and the method of choosing, for each MB, whether or not to use that long-term reference frame (or use the short-term or intra mode) can work together synergistically or can oppose each other. The gains in performance ranged from 0.6 to 1.6 dB.

The results point to the superiority of HQ (pulsed quality) over regular updating for lossy packet network video transmission with a dual-frame coder. This gain comes at trivial extra computational and implementation cost and can be easily deployed in a standard compliant H.264 codec.

In our work, the heuristic allocation we used worked remarkably well for all combinations of image sequences and bit rates, but there is still much that has not been modeled and optimized. Future work will concentrate on finding an efficient explicit rate control mechanism to allocate rate to the long-term and regular frames. Finding good update parameters is at least as challenging.

REFERENCES

[1] M. Gothe and J. Vaisey, "Improving motion compensation using multiple temporal frames," in *Proc. IEEE Pacific Rim Conf. Communications, Computers, and Signal Processing*, vol. 1, May 1993, pp. 157–160.

[2] N. Vasconcelos and A. Lippman, "Library-based image coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. v, Apr. 1994, pp. 489–492.

[3] T. Wiegand, X. Zhang, and B. Girod, "Long-term memory motion-compensated prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 70–84, Feb. 1999.

[4] T. Wiegand, Final draft international standard for joint video specification H.264, in JVT of ISO/IEC MPEG and ITU-T VCEG, JVT-G050, Mar. 2003.

[5] T. Fukuhara, K. Asai, and T. Murakami, "Very low bit-rate video coding with block partitioning and adaptive selection of two time-differential frame memories," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 212–220, Feb. 1997.

[6] R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE J. Select. Areas Commun.*, vol. 18, pp. 966–976, June 2000.

[7] A. Leontaris and P. C. Cosman, "Video compression with intra/inter mode switching and a dual frame buffer," in *Proc. IEEE Data Compression Conf.*, Mar. 2003, pp. 63–72.

[8] V. Chellappa, P. C. Cosman, and G. M. Voelker, "Dual frame motion compensation with uneven quality assignment," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 2004, pp. 262–271.

[9] A. Leontaris and P. C. Cosman, "Video compression for lossy packet networks with mode switching and a dual-frame buffer," *IEEE Trans. Image Processing*, vol. 13, pp. 885–897, July 2004.