RATE ALLOCATION ALGORITHMS FOR MOTION COMPENSATED EMBEDDED VIDEO CODERS

Josep Prades-Nebot[†], *Gregory W. Cook*[‡] *and Edward J. Delp*[*]

[†] Departamento de Comunicaciones-iTEAM
Universidad Politécnica de Valencia
Valencia 46071, SPAIN
jprades@dcom.upv.es

[‡] Thomson
Corporate Research
Indianapolis, IN 46202, USA
greg.cook@thomson.net

[*]Video and Image Processing Laboratory (VIPER)
Purdue University
West Lafayette, IN 47907-1285, USA
ace@purdue.ecn.edu

ABSTRACT

In this paper, we present two rate allocation algorithms for embedded motion compensated video coders. The algorithms are based on the modeling of both the video signal and the coder which allow us to express the coding distortion with a recurrence equation. Our algorithms assign rates to the frames of each Group of Pictures (GOP) of a video sequence in an optimum way. In the first algorithm, the criterion is to minimize the average (MINAVE) distortion and in the second to achieve constant distortion (CD) in all frames. Numerical simulations show the MINAVE criterion can introduce large variations in quality with no significant gains in average distortion with respect to the CD criterion. We also show how the the motion estimation accuracy and the GOP length influence in both strategies.

1. INTRODUCTION

In embedded coders, the rate can be set and changed in an easy and precise way [1–3]. This property allows these kinds of coders to adapt to changes in the available bandwidth and to different receiver capabilities. Furthermore, directly controlling the rate [4] is not necessary and just rate allocation (RA) algorithms are required to distribute the available bits between the different coding units (*e.g.,* frames and macroblocks). Thanks to these advantages, embedded coding has been included in the image compression standard JPEG2000 [3] and in the video streaming profile of the standard MPEG-4 [5]. In this paper, we present two bit allocation algorithms for Motion Compensated Embedded (MCE) video coders.

Two main approaches exist for the RA problem: algorithms based on models [6] and algorithms based on operational rate-distortion (R-D) optimization [7]. Rate allocation algorithms based on models can provide closed-form expressions and have a low computational cost, but the allocations provided can be far from the optimum ones because it is difficult to model video sources and coders accurately. On the other hand, algorithms based on operational R-D optimization provide exact optimum solutions for each particular signal and coding algorithm but at the expense of a higher computational cost. This computational complexity is especially high when they are used to optimize MCE video coders, due to the temporal dependency between frames and the very large number of points on the R-D curve of embedded coders.

In this paper, we present two RA algorithms for MCE video coders assuming models for both the video sequences and the motion compensated prediction operation of the video coder. We assume that the coding algorithm segments the video sequence in groups of pictures (GOP) which are coded independently. The first RA algorithm provides a minimum average (MINAVE) distortion in the frames of the video sequence. As large quality variations in the frames of the video sequence are annoying visually, some RA algorithms try to achieve constant distortion (CD) instead of minimizing the average distortion [8]. Our second RA algorithm provide constant distortion in all frames in the GOP.

The paper is organized as follows. In Section 2, we describe the models used and the hypotheses assumed in our analysis. In Section 3, we analyze the efficiency of the embedded MCP-based video coder assuming input video signals are partitioned in GOPs of equal length. In Sections 4 and 5, we present the minimum average distortion and the constant quality RA algorithms respectively. In Section 6, we show numerical results of both algorithms for different motion estimation accuracy and GOP lengths. Finally, in Section 7 we summarize our results and comment on future work.

2. MODELS AND HYPOTHESES ASSUMED

In the following, x and y are the spatial variables, and t is the temporal variable of the digital video sequence, with $x, y, t \in \mathbb{Z}$. Their corresponding frequency variables are ω_x, ω_y and ω_t respectively. For simplicity, sometimes we use $\Lambda = (\omega_x, \omega_y)$ and since in our study the signals and systems involved are discrete, when specifying spatial spectra of signals or frequency responses of systems we only specify them in the base-band, that is, in $\Lambda_B = \{(\omega_x, \omega_y) | -\pi \leq \omega_x, \omega_y \leq \pi\}$. When there is no loss in clarity, some signal variables or even all of them are removed.

In our theoretical analysis we make some assumptions about the signals and systems involved. The quantization noise $q[t]$ introduced in the intra-frame encoding of the t-th frame is modeled as an additive zero-mean white noise whose variance $\sigma_q^2[t]$ is

$$\sigma_q^2[t] = \sigma_e^2[t] \; 2^{-\beta_t R_t} \tag{1}$$

where $\sigma_e^2[t]$ is the power of the predicted error frame $e[t]$, R_t is the rate used to encode the t-th frame and β_t is a parameter that measures the efficiency of the intra-frame coding of the t-th frame [6]. The value of parameter β depends on frame content and on the type (intra/inter) of frame encoding. We assume the quantization noise is uncorrelated with the signal that is being quantized, which is approximately valid for large R.

This work has been supported by a Generalitat Valenciana grant (Project GV04B-478) and by the program CICYT TIC-2002-02469.

The rest of assumptions are similar to the ones assumed in [9]. With respect to the input video signal s, we assume that its frames constitute a stationary discrete random field and that the only difference between consecutive frames is a constant-in-time and uniform-in-space displacement (d_x,d_y). The predictor is modeled as a random linear time-invariant system whose frequency response is

$$H(\omega_x, \omega_y, \omega_t) = F(\omega_x, \omega_y)\, e^{-j(\omega_x \hat{d}_x + \omega_y \hat{d}_y + \omega_t)} \qquad (2)$$

where $F(\omega_x, \omega_y)$ is the frequency response of a spatial filter and ($\hat{d}_x$,$\hat{d}_y$) is the estimated (random) displacement vector. In general, in the motion estimation there is a random displacement error vector $\Delta d = (\Delta d_x, \Delta d_y)$, where

$$(\Delta d_x, \Delta d_y) = (d_x, d_y) \; - \; (\hat{d}_x, \hat{d}_y) \qquad (3)$$

with a probability density function $p_{\Delta d}(\Delta d)$.

Although in real video sequences the hypothesis of a constant-in-time and uniform-in-space translational motion is not accurate, it allows study of the rate allocation problem and demonstrates the influence of different factors, *e.g.*, motion estimation accuracy, intra-frame coding efficiency, and GOP length. In our analysis, the rate needed to encode the motion vectors is ignored; this does not have an important influence except when encoding at very low rates.

3. ANALYSIS OF THE MCE VIDEO CODER

Figure 1 shows the block diagram used to analyze the efficiency of a motion compensated embedded video coder. The input of the transmitter is a monochrome discrete stationary video signal $s[t]$ of N frames ($0 \leq t \leq N-1$). We assume the first frame ($t = 0$) is intra-coded (I-frame) and the rest ($1 \leq t \leq N-1$) are inter-coded with previous-frame-based prediction (P-frames). In Figure 1, intra or inter coding depends on the switches' state (intra if open, inter if closed). In every instant of time t, the difference between the original frame $s[t]$ and a prediction of it $\hat{s}[t]$ is computed, generating a sequence of prediction error frames (PEF) $e[t]$. We assume the prediction of frame at time t is only based on the previous frame ($t-1$). Prediction is based on a reconstructed frame $s'[t]$, which is a version of $s[t]$ with quantization noise ($q[t]$). Motion Estimation (ME) is performed in order to estimate local displacements between every pair of consecutive frames. In this way, the encoder can make a motion compensated prediction (MCP) of every input frame, which is represented in Figure 1 by the filter H. Each PEF $e[t]$ is encoded using an intra-frame encoder that performs transform, quantization and entropy coding. Each compressed frame is then decoded (entropy decoded, dequantized and inverse transformed) in order to provide a distorted version of the PEF ($e'[t]$) to the MCP loop of the encoder.

Thanks to the closed-loop structure of the MCE coder

$$r[t] = s''[t] - s[t] = q[t] \qquad (4)$$

and consequently the distortion D_t of the t-th frame is

$$D_t = \sigma_r^2[t] = \sigma_q^2[t], \quad 0 \leq t \leq N-1. \qquad (5)$$

For a rate R, the rate allocation algorithm assigns a rate R_t to encode each PEF $e[t]$ (with a parameter β_t).

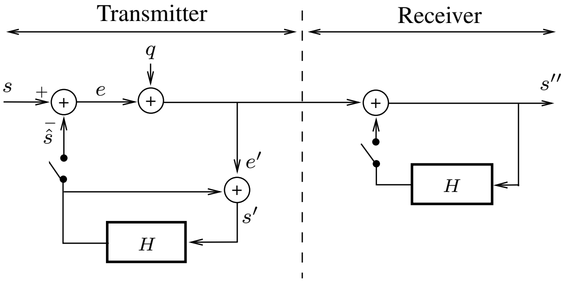

Fig. 1. Block diagram to analyze the MCE video coder.

The distortion of the I-frame is

$$D_0 = \sigma_s^2\, 2^{-\beta_0\, R_0} \qquad (6)$$

where σ_s^2 is the power of the video signal. For the rest of frames, the prediction error frame at time t can be expressed as

$$\begin{aligned} e[x,y,t] &= s[x,y,t] - s[x,y,t-1] * h[x,y] \\ &+ \; q[x,y,t-1] * h[x,y], \quad 1 \leq t \leq N-1, \end{aligned} \qquad (7)$$

where "$*$" denotes spatial convolution and $h[x,y]$ is the impulse response of a system that involves the spatial filtering ($f[x,y]$) and the spatial displacement ($\hat{d}_x, \hat{d}_y$). From (7), and after some manipulations [10], we obtain

$$\begin{aligned} S_{ee}(\Lambda, t) &= S_{ss}(\Lambda)\, \left[1 - 2\,\mathrm{Re}\left\{F(\Lambda)P^*(\Lambda)\right\} + |F(\Lambda)|^2\right] \\ &+ \; |F(\Lambda)|^2\, S_{qq}(\Lambda, t-1) \end{aligned} \qquad (8)$$

where $P(\Lambda)$ is the Fourier Transform of $p_{\Delta d}(\Delta d)$. After integration of (8) in Λ_{B} we obtain

$$\sigma_e^2[t] = E_s + D_{t-1} E_f \qquad (9)$$

where $\sigma_e^2[t]$ is the PEF variance at time t, E_s is

$$E_s = \frac{1}{4\pi^2} \iint_{\Lambda_B} S_{ss}(\Lambda)\, \left[1 - 2\,\mathrm{Re}\left\{F(\Lambda)\, P^*(\Lambda)\right\} + |F(\Lambda)|^2\right]\, d\Lambda,$$

and E_f is

$$E_f = \frac{1}{4\pi^2} \iint_{\Lambda_B} |F(\Lambda)|^2\, d\Lambda. \qquad (10)$$

From (1) and (9) we obtain

$$D_t\, 2^{\beta_t\, R_t} = E_s + D_{t-1} E_f, \quad 1 \leq t \leq N-1 \qquad (11)$$

which allows us to obtain D_t recursively with (6) being the initial value of the recurrence.

4. RATE ALLOCATION FOR MINIMUM MEAN DISTORTION

The RA problem providing minimum average distortion can be formulated as:

$$\text{minimize} \quad D = \frac{1}{N} \sum_{t=0}^{N-1} D_t \qquad (12)$$

$$\text{subject to} \quad \frac{1}{N} \sum_{t=0}^{N-1} R_t = R. \qquad (13)$$

By using the Lagrange multiplier method, the constrained optimization problem in (12) and (13) can be transformed into an unconstrained optimization problem, where the optimum RA strategy is the set of rates minimizing the Lagrangian cost function

$$J(R_0,\dots,R_{N-1}) = \frac{1}{N}\sum_{t=0}^{N-1} D_t + \lambda\left[\frac{1}{N}\sum_{t=0}^{N-1} R_t \,-\, R\right]. \tag{14}$$

To solve this MINAVE optimization problem, an algorithm similar to the one described in [6] is used. In fact, we also use the same R-D model to characterize the intra-frame encoding process that in [6]. The main difference between our MINAVE algorithm and the one present in [6] is how the frame dependency problem is taken into account. In [6], the solution is mainly based on the assumption that the variance of the motion compensated residue is an affine function of the reference frame distortion through a parameter α. In our algorithm, the models and hypotheses described in Section 2 allows us to obtain explicitly the dependency through (11).

The optimum RA is achieved when

$$\frac{\partial J(R_0,\dots,R_{N-1})}{\partial R_t} = 0, \quad 0 \le t \le N-1, \tag{15}$$

which, due to the temporal dependence between frames, provides

$$\frac{\partial J}{\partial R_t} = \frac{1}{N}\left[\frac{\partial D_t}{\partial R_t} + \frac{\partial D_{t+1}}{\partial R_t} + \cdots + \frac{\partial D_{N-1}}{\partial R_t}\right] + \frac{\lambda}{N} = 0 \tag{16}$$

for $0 \le t \le N-1$. For the last frame ($t = N-1$), we have

$$\frac{\partial D_{N-1}}{\partial R_{N-1}} + \lambda = 0 \tag{17}$$

which provides

$$D_{N-1} = \frac{\lambda}{\beta_{N-1}\ln 2}. \tag{18}$$

If $t = N-2$ in (16), we have

$$\left[\frac{\partial D_{N-2}}{\partial R_{N-2}} + \frac{\partial D_{N-1}}{\partial R_{N-2}}\right] + \lambda = 0, \tag{19}$$

which provides

$$D_{N-2} = \frac{\lambda}{\beta_{N-2}\ln 2\,\left(1 + E_f\, 2^{-\beta_{N-1}R_{N-1}}\right)}. \tag{20}$$

Additionally, from (11) and (20), we obtain

$$D_{N-2}^2 + \left[\frac{E_s}{E_f} - \frac{\lambda}{\beta_{N-2}\ln 2} + D_{N-1}\right] D_{N-2} - \frac{\lambda\, E_s}{\beta_{N-2}\ln 2\, E_f} = 0, \tag{21}$$

whose positive solution provides D_{N-2}.

For other instants of time different to $N-1$ and $N-2$, the following identity is useful

$$\frac{\partial D_{t+k}}{\partial R_t} = \frac{\partial D_t}{\partial R_t}\, E_f^k \prod_{l=1}^{k} 2^{-\beta_{t+l}R_{t+l}}, \tag{22}$$

for $0 \le t \le N-2$ and $1 \le k \le N-1-t$. If we define

$$P_t \triangleq 1 + \sum_{k=1}^{N-1-t} E_f^k \prod_{l=1}^{k} 2^{-\beta_{t+l}R_{t+l}}, \quad 0 \le t \le N-2 \tag{23}$$

and $P_{N-1} \triangleq 1$, then (16) transforms into

$$D_t\, \beta_t \ln 2\, P_t = \lambda, \quad 0 \le t \le N-1. \tag{24}$$

By taking into account

$$P_t = 1 + E_f\, 2^{-\beta_{t+1}R_{t+1}}\, P_{t+1}, \quad 0 \le t \le N-2 \tag{25}$$

Equation 24 can be written as

$$D_t\, \beta_t \ln 2\, \left(1 + E_f\, 2^{-\beta_{t+1}R_{t+1}}\, P_{t+1}\right) = \lambda, \tag{26}$$

for $0 \le t \le N-2$. This last equation together with (11) provides

$$D_t^2 + \left[\frac{E_s}{E_f} - \frac{\lambda}{\beta_t \ln 2} + D_{t+1}P_{t+1}\right] D_t - \frac{\lambda\, E_s}{\beta_t \ln 2\, E_f} = 0, \tag{27}$$

whose solution allow us to get D_t from D_{t+1} and P_{t+1} for $0 \le t \le N-2$. In fact, if we define

$$B_t \triangleq \frac{E_s}{E_f} - \frac{\lambda}{\beta_t \ln 2} + D_{t+1}P_{t+1} \tag{28}$$

$$C_t \triangleq \frac{-\lambda E_s}{\beta_t \ln 2\, E_f} \tag{29}$$

then D_t can be computed trough

$$D_t = \frac{-B_t + \sqrt{B_t^2 - 4C_t}}{2}, \tag{30}$$

for $0 \le t \le N-2$ (D_{N-1} can be computed from (18)).

With respect to the rates R_t, from (11) we obtain

$$R_{t+1} = \frac{1}{\beta_{t+1}} \log_2 \frac{E_s + D_t E_f}{D_{t+1}}, \quad t = 0,\dots,N-2. \tag{31}$$

Then by using (25), (30) and (31) recursively, we can get the rates of all frames for a value of λ. The recursion starts with D_{N-1}, whose value is obtained from (18), and with $P_{N-1} = 1$. Then, by using (25), (30) and (31) recursively we can get D_t in $t = N-2,\dots,0$ and R_t in $t = N-1,\dots,1$. Finally, R_0 can be obtained through

$$R_0 = \frac{1}{\beta_0} \log_2 \frac{\sigma_s^2}{D_0}. \tag{32}$$

The algorithm should be repeated for several λ values until the rate constraint is met. Thanks to the convexity property of the rate-distortion function, the bisection algorithm can be used.

It is straightforward to determine that if all P-frames have the same β value ($\beta_t = \beta_{\mathrm{P}}$ in $1 \le t \le N-1$), then D_t has the same value in $1 \le t \le N-2$ and R_t has the same value in $2 \le t \le N-2$. It is interesting to notice that, while assigning the same number of bits for all P-frames in a GOP is usually used as a non-optimum but simple RA strategy, when β is equal for all P-frames, the MINAVE criterion assigns the same rate to all P-frames except the first and the last one.

5. RATE ALLOCATION FOR CONSTANT DISTORTION

In this section we study the rate allocation (RA) problem for achieving constant distortion (CD). If all the frames have the same distortion, from (6) and (11) we get

$$\sigma_s^2\, 2^{-\beta_0 R_0} = \frac{E_s}{2^{\beta_t R_t} - E_f}, \quad 1 \le t \le N-1. \tag{33}$$

which together with the rate constraint (13) constitutes a system of equations whose solution is the RA providing constant distortion. If $\beta_t = \beta_{\text{P}}$ in $1 \leq t \leq N-1$, all P-frames must have the same rate

$$R_t = R_{\text{P}}, \quad 1 \leq t \leq N-1, \tag{34}$$

and from (33) together with the rate constraint

$$RN = R_0 + (N-1)R_{\text{P}} \tag{35}$$

we get

$$\frac{E_s}{2^{\beta_{\text{P}}\, R_{\text{P}}} - E_f} = \sigma_s^2\, 2^{-\beta_0 [RN-(N-1)R_P]} \tag{36}$$

whose solution is the optimum R_{P} (optimum R_0 can be obtained later from (35)).

6. NUMERICAL SIMULATIONS

In this section we show the results of numerical simulations obtained when using the rate allocation (RA) algorithms of Sections 4 and 5 and the following settings. Similarly to [9], the input signal is the discrete random field (DRF) that results after an ideal low-pass filtering and sampling of a continuous random field with isotropic autocorrelation and one-step (in x and y) correlation coefficients ρ equal to 0.93. The variance of the DRF is $\sigma_s^2 = 0.98$. We assume that Δd follows a zero mean, Gaussian isotropic probability density function with variance $\sigma_{\Delta d}^2 = 0.2\,T^2$ where T is the spatial sampling period. The rate is $R = 0.493$ bits/pixel (which for video in CIF format corresponds to 1.5 Mb/s), $N = 15$, $F(\Lambda) = 1$, $\beta_0 = 6$ and $\beta_t = 3$ for $1 \leq t \leq N-1$.

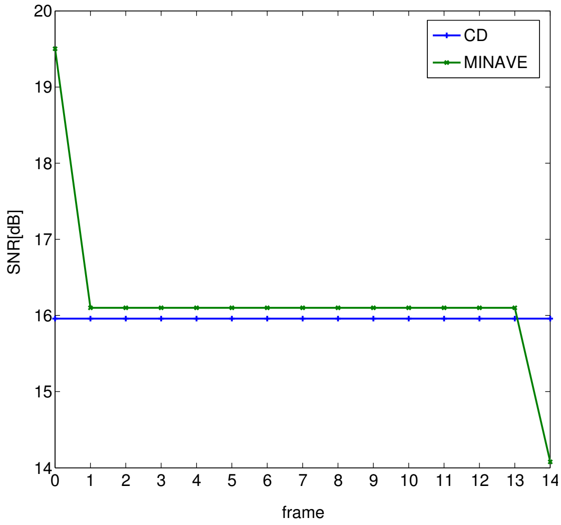

Fig. 2. SNR as a function of the frame number using the MINAVE and CD criteria.

Figure 2 shows the SNR as a function of the frame number obtained with the previously described settings for the MINAVE and the CD strategies. Notice that the MINAVE strategy provides large SNR variations in the two first and two last frames of the sequence. Consequently, large quality fluctuations can be introduced in the transitions between GOPs if the MINAVE strategy is used. The MINAVE strategy, however, does not provide a significant gain in average SNR with respect to the CD strategy (16.19 dB with MINAVE and 15.96 dB with CD).

The gain in average SNR of the MINAVE strategy with respect to the CD one does not vary significantly when we vary the motion estimation accuracy (+0.254 dB if $\sigma_{\Delta d}^2 = 0.05\,T^2$, +0.233 dB if $\sigma_{\Delta d}^2 = 0.2\,T^2$ and +0.231 dB if $\sigma_{\Delta d}^2 = T^2$). With respect to the GOP length, the shorter the length, the larger the gain in mean SNR of the MINAVE strategy with respect to the CD (+0.79 dB with $N = 5$, +0.233 dB with $N = 15$ and +0.04 dB with $N = 100$).

7. CONCLUSION AND FUTURE WORK

In this work, we have presented two rate allocation algorithms (MINAVE and CD) for embedded video coders using motion compensated prediction. Numerical simulations shows that the MINAVE criterion can introduce large variations in quality and only provide significant improvement in average distortion with short GOP lengths.

8. REFERENCES

[1] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3345–3462, Dec. 1993.

[2] A. Said and W. A. Pearlman, "New, fast and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 3, pp. 243–250, June 1996.

[3] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice*. Kluwer Academic Publishers, 2002.

[4] Test Model 5 (TM5), ISO/IEC JTC1/SC29/WG11/93-225b Test Model Editing Committee, Apr. 1993.

[5] Information technology–Coding of audio-visual objects–Part 2: Visual Amendment 4: Streaming Video Profile, MPEG 2000/N3518, 2000.

[6] P.-Y. Cheng, J. Li, and C.-C. J. Kuo, "Rate control for an embedded wavelet video coder," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 4, pp. 696–701, August 1997.

[7] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Mag.*, vol. 15, no. 6, pp. 23–50, November 1998.

[8] G. M. Schuster, G. Melnikov, and A. Katsaggelos, "A review of the minimum maximum criterion for optimal bit allocation among dependent quantizers," *IEEE Trans. Multimedia*, vol. 1, no. 1, pp. 3–17, March 1999.

[9] B. Girod, "Motion-compensating prediction with fractional-pel accuracy," *IEEE Trans. Commun.*, vol. 41, no. 4, pp. 604–611, Apr. 1993.

[10] J. Prades-Nebot, G. W. Cook, and E. J. Delp, "An analysis of the efficiency of different SNR-scalable strategies for video coders," *IEEE Trans. Image Processing (accepted for publication)*.