ANALYSIS OF CODING EFFICIENCY OF MOTION-COMPENSATED INTERPOLATION AT THE DECODER IN DISTRIBUTED VIDEO CODING

M. Tagliasacchi, L. Frigerio, S. Tubaro

Dipartimento di Elettronica e Informazione
Politecnico di Milano,
Milan - Italy

ABSTRACT

This paper analyzes the coding efficiency of distributed video coding (DVC) schemes that perform motion-compensated interpolation at the decoder. The decoder has access only to the key frames when generating the side information for intermediate frames. This fact introduces a displacement estimation error that depends on several factors: 1) the overall motion complexity; 2) the temporal coherence of the motion field; 3) the temporal distance between successive key frames. Adopting a state-space model and a Kalman filtering framework, we obtain an estimate of the displacement error variance. This is used to determine the rate-distortion function of the overall coding scheme, that takes into account both intra-coded key frames and DVC-coded frames. The proposed model shows that motion-compensated interpolation is unable to achieve the coding efficiency of conventional motion-compensated predictive coding.

Index Terms— Video coding, motion analysis, distributed video coding

1. INTRODUCTION

Distributed Video Coding (DVC) is a recent video coding paradigm whose main idea is to perform intra-frame encoding and inter-frame decoding. Results obtained on test video sequences reveal that DVC coding schemes generally improve the coding efficiency with respect to intra-frame coding, but, so far, they have been unable to achieve the coding efficiency of conventional motion-compensated predictive codecs, at least for the case of noise free transmission [1].

The goal of this paper is to introduce a model that allows to study the coding efficiency of DVC-based coding schemes. We restrict our analysis to schemes that compute the side information at the decoder by performing motion-compensated interpolation, starting from two intra-coded key frames [1]. Specifically, we focus only on the generation of the side information, neglecting other factors related to the channel coding tools that are typically used to replace conventional entropy coding. We elaborate our model in two steps. First, for each Wyner-Ziv coded frame, we estimate the displacement error variance introduced by motion-compensated interpolation. In fact, the true motion field is not directly available at the decoder, and it must be estimated introducing displacement estimation errors. Then, we estimate the power spectral density of the motion-compensated prediction error to obtain the rate-distortion curves by inverse water-filling [2]. Armed with the proposed model, we investigate the trade-offs between motion-compensated interpolation accuracy and GOP size, in order to find the optimal GOP size for a target distortion.

This paper extends our previous work in [3] in two ways: arbitrary GOP lengths are considered and the analysis is not restricted to high rates, thus including the effect of lossy key frames. In addition, experimental results on real video sequences are presented to corroborate the validity of the proposed model. A similar work appeared in [4], where the model explicitly addresses only the case of motion-extrapolation.

2. RATE-DISTORTION MODEL

Consider a GOP of size N frames, encoded either using a conventional motion-compensated predictive codec or a DVC-based scheme as in [1]. These schemes differ in the way the motion-compensated prediction (side information) $\hat{s}(t)$ of the current frame $s(t)$ is generated:

- *Motion estimation at the encoder*: $\hat{s}(t) = \hat{s}_P(t)$ is obtained by exploiting data from the current frame $s(t)$ and from the previously encoded frames $s'(t-1)$ (s' is the quantized version of s). An $I-P-P-\ldots-I$ GOP structure is assumed.
- *Motion-compensated interpolation at the decoder*: $\hat{s}(t) = \hat{s}_{WZ}(t)$ is generated at the decoder side. The current frame is not available. The decoder performs motion interpolation using lossy coded key frames $s'(\tau_1)$ and $s'(\tau_2)$ only ($\tau_1 < t < \tau_2$) [5][6]. An $I-WZ-WZ-\ldots-I$ GOP is adopted, where the decoding of any Wyner-Ziv (WZ) frame requires both the previous and the next I frames to be decoded first.

If we constrain the distortion D to be constant along the GOP, the average rate R per frame can be computed as:

$$R(D) = \frac{1}{N}\left[R^I(D) + \sum_{i=1}^{N-1} R_i^{\{P,WZ\}}(D)\right], \qquad (1)$$

where $R^I(D)$ is the contribution of the intra-coded frame and $R_i^{\{P,WZ\}}(D)$ that of the ith inter-coded frame (for the case of motion-compensated prediction at the encoder or motion-compensated interpolation at the decoder).

The rate-distortion curve $R^I(D)$ is given by the following parametric set of equations [7]:

$$D^I(\theta) = E[(s'-s)^2] = \frac{1}{4\pi^2}\iint_\Lambda \min[\theta, \phi_{ss}(\Lambda)]d\Lambda \qquad (2)$$

$$R^I(\theta) = \frac{1}{8\pi^2}\iint_\Lambda \max\left[0, \log_2\frac{\phi_{ss}(\Lambda)}{\theta}\right]d\Lambda \quad \text{bit}, \qquad (3)$$

where $\phi_{ss}(\Lambda)$ ($\Lambda = (\omega_x, \omega_y)$) is the spatial power spectral density (PSD) of the source and $\theta > 0$ is a real-valued parameter that allows to move along the rate-distortion curve. The latter is proportional to the amount of distortion introduced by quantization.

In the following, we derive the curves $R^{\{P,WZ\}}(D)$ adopting the framework introduced in [2]. To this end, let us denote the residual frame after motion-compensated prediction as $e(t) = s(t) - \hat{s}(t)$ and define the spatial power spectral density of $e(t)$ as $\phi_{ee}(\Lambda)$. Let us consider a video signal that is described by a constant, translatory displacement (d_x, d_y), and neglect any other effect like rotation, zoom, occlusions, illumination changes, etc. The approximate expression of $\phi_{ee}(\Lambda)$ is given by [2]

$$\phi_{ee}(\Lambda) \approx \begin{cases} \phi_{ss}(\Lambda) & \text{if } \phi_{ss}(\Lambda) < \theta \\ \max\{\tilde{\phi}_{ee}(\Lambda), \theta\} & \text{otherwise} \end{cases},$$

$$\tilde{\phi}_{ee}(\Lambda) = 2\phi_{ss}(\Lambda)(1 - e^{-\frac{1}{2}(\omega_x \sigma^2_{\Delta d_x} + \omega_y \sigma^2_{\Delta d_y})}) + \theta, \quad (4)$$

where $\sigma^2_{\Delta d_c}$ denotes the variance of the displacement error $\Delta d_c = d_c - \hat{d}_c$ $(c = x, y)$, which is assumed to be zero mean and Gaussian distributed. The error is strictly connected to the way motion is estimated and represented, as it will be detailed shortly.

In [2], an approximation of the rate-distortion function is given by

$$D^{\{P,WZ\}}(\theta) = E[(e' - e)^2] = \frac{1}{4\pi^2} \iint_{\Lambda} \min[\theta, \phi_{ss}(\Lambda))] d\Lambda \quad (5)$$

$$R^{\{P,WZ\}}(\theta) = \frac{1}{8\pi^2} \iint_{\Lambda:(\phi_{ss}(\Lambda)>\theta \text{ and } \tilde{\phi}_{ee}(\Lambda)>\theta)} \log_2 \frac{\tilde{\phi}_{ee}(\Lambda)}{\theta} d\Lambda \quad (6)$$

We can observe that, in order to compute equation (1), we need to characterize the values of the displacement error variances $\sigma^2_{\Delta d_x}$ and $\sigma^2_{\Delta d_y}$ for each frame in the GOP. Assuming isotropic displacement errors, we can state that, on average, $\sigma^2_{\Delta d_x} = \sigma^2_{\Delta d_y} = \sigma^2_{\Delta d}$. Therefore we will drop the coordinate index x, y in the rest of this paper. We can analyze the following two cases:

- P frames: The motion estimation is performed at the encoder. We can assume that the displacement error is solely due to the finite accuracy used to represent motion vectors ($M = 1, 1/2, 1/4, \ldots$ pixels). Therefore, we can write $\sigma^2_{\Delta d} = M^2/12$ for any frame in the GOP as indicated in [2].
- WZ frames: The motion estimation is performed at the decoder between successive intra-coded key frames. Then, this is used to infer the motion for intermediate WZ frames. In order to evaluate $\sigma^2_{\Delta d_i}$ for the ith frame we propose a model based on Kalman filtering, detailed in the following section.

3. STATE-SPACE MOTION MODEL

In this section, we introduce a state-space model according to the Kalman filtering framework. We describe the time evolution of the true displacements with the state equation, and the noisy observation of the motion between two intra-coded key frames with the output equation.

Specifically, we introduce the following state equation

$$d(t) = \rho d(t-1) + z(t) \quad (7)$$

where $d(t)$ is the true displacement that the frame $s(t)$ is subject to, ρ is the temporal correlation coefficient and $z(t)$ ia a zero-mean white noise, having variance σ^2_z. The variance of $d(t)$ can be computed as $\sigma^2_d = \sigma^2_z/(1-\rho^2)$. In order to gain an insight, we can interpret σ^2_d as an indication of the motion complexity; a high value of σ^2_d suggests that large displacements are expected. On the other hand, ρ measures the temporal coherence of the motion field, for a given value of σ^2_d.

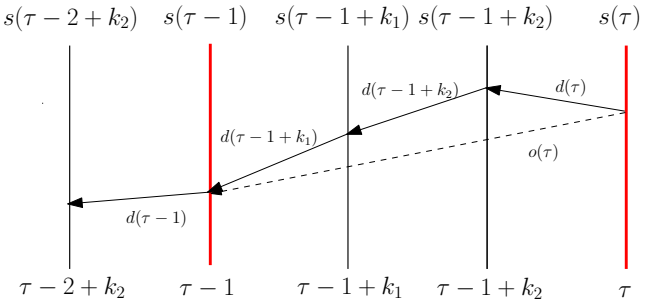

Fig. 1. Motion-compensated interpolation with time step τ referred to the evolution of the intra-coded key frames

A value of ρ close to one indicates that motion has approximately uniform velocity along time.

In the proposed model, we can view the motion-compensated interpolation process as an estimation of the displacements at time $t, t-1, \ldots, t-N+1$ (i.e. $\hat{d}(t), \hat{d}(t-1), \ldots, \hat{d}(t-N+1)$), when only the motion $o(t)$ between two key frames is observed.

$$o(t) = d(t) + d(t-1) + d(t-2) + \ldots + d(t-N+1) + w(t) \quad (8)$$

where $w(t)$ is a white noise $WN(0, \sigma^2_w)$ that takes into account the finite accuracy of displacements ($\sigma^2_w = M^2/12$), as already explained for P frames in the previous section.

The state-space model described by equation (7) and (8), implies that a new observation $o(t)$ is available at any time instant t. Actually, we have access only to one observation every N time instants, where N is the GOP size. A more accurate model for the problem at hand is obtained by relating the increment of the time variable to intra-frames only. With a change of variables, we define $\tau = t/N$ and we rewrite the state-space model in the new time units τ.

For the sake of simplicity, consider a GOP of $N = 3$ frames (see Figure 1). At time τ the intra-frames $s(\tau)$ and $s(\tau - 1)$ are used to compute the displacement $o(\tau)$. WZ frames are defined at intermediate fractional times $\tau - 1 + k_1$ and $\tau - 1 + k_2$ ($k_i = i/N$). Exploiting the autoregressive model (7) and denoting $d_i(\tau) = d(\tau - 1 + k_i)$ and $z_i(\tau) = z(\tau - 1 + k_i)$ we obtain the subsequent model:

$$\begin{aligned} d_1(\tau) &= \rho d(\tau-1) + z_1(\tau) \\ d_2(\tau) &= \rho^2 d(\tau-1) + \rho z_1(\tau) + z_2(\tau) \\ d(\tau) &= \rho^3 d(\tau-1) + \rho^2 z_1(\tau) + \rho z_2(\tau) + z(\tau) \\ o(\tau) &= d_1(\tau) + d_2(\tau) + d(\tau) + w(\tau) \end{aligned} \quad (9)$$

that can be written in the canonical form prescribed by Kalman filtering:

$$\mathbf{d}(\tau) = F\mathbf{d}(\tau-1) + \mathbf{v}_1(\tau) \quad (10)$$

$$o(\tau) = H\mathbf{d}(\tau) + v_2(\tau) \quad (11)$$

where $\mathbf{d}(\tau) = [d_1(\tau),\ d_2(\tau),\ d(\tau)]^T$, $\mathbf{v}_1(\tau) = [z_1(\tau),\ \rho z_1(\tau) + z_2(\tau),\ \rho^2 z_1(\tau) + \rho z_2(\tau) + z(\tau)]^T$, $v_2(\tau) = w(\tau)$.

For a GOP of size N, we can generalize the previous discussion and we obtain the following matrices:

$$F_{(N\times N)} = \begin{pmatrix} \rho & 0 & \cdots & 0 \\ \rho^2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \rho^N & 0 & \cdots & 0 \end{pmatrix} \quad (12)$$

$$H_{(1\times N)} = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} \quad (13)$$

$$V_{2(1\times 1)} = E[v_2^2(\tau)] = \sigma_w^2 \quad (14)$$

$$V_{1(N\times N)} = E[\mathbf{v}_1(\tau)\mathbf{v}_1(\tau)^T] = \quad (15)$$
$$\sigma_z^2 \begin{pmatrix} 1 & \rho & \cdots & \rho^{N-1} \\ \rho & \rho^2+1 & \cdots & \rho^N+\rho^{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{N-1} & \rho^N+\rho^{N-2} & \cdots & \sum_{i=1}^{N}\rho^{2(N-i)} \end{pmatrix}$$

The matrix V_{12} is composed of zeros, because the noise terms $\mathbf{v}_1(\tau)$ and $v_2(\tau)$ are uncorrelated.

Going back to our original problem, we want to obtain the variances of the displacement errors $\sigma^2_{\Delta d_i}$ of the ith WZ frame in the GOP. Let us consider $\hat{\mathbf{d}}(\tau|\tau-1)$, i.e. the estimation of the state vector $\mathbf{d}(\tau)$ computed at time τ with data available up to time $\tau-1$. Kalman theory states that it is possible to relate the variance of the error on the state of the Kalman *predictor* ($\Delta\mathbf{d}(\tau|\tau-1) = \mathbf{d}(\tau) - \hat{\mathbf{d}}(\tau|\tau-1)$) at time τ with that at time $\tau-1$ via the RDE (Riccati Differential Equation):

$$P(\tau+1) = FP(\tau)F^T + V_1 - K(\tau)(HP(\tau)H^T + V_2)K^T \quad (16)$$

where $P(\tau) = E[\Delta\mathbf{d}(\tau|\tau-1)\Delta\mathbf{d}^T(\tau|\tau-1)]$ and the Kalman gain $K(\tau)$ is defined as $K(\tau) = (FP(\tau)H^T + V_{12})(HP(\tau)H^T + V_2)^{-1}$. When the observation at time τ is available, in addition to those up to time $\tau-1$, the variance of the error on the state ($\Delta\mathbf{d}(\tau|\tau) = \mathbf{d}(\tau) - \hat{\mathbf{d}}(\tau|\tau)$) of the Kalman *filter* must be considered, instead of the one of the Kalman *predictor*:

$$E[\Delta\mathbf{d}(\tau|\tau)\Delta\mathbf{d}^T(\tau|\tau)] = \quad (17)$$
$$P_{filt}(\tau) = P(\tau) - P(\tau)[H^T[HP(\tau)H^T] + V_2]^{-1}HP(\tau)$$

In (16), upon convergence, $P(\tau+1) = P(\tau) = P$. Substituting P into eq. (16), we obtain the ARE (Algebrical Riccati Equation) and we solve by P. Values of the matrix $P_{filt}(\tau)$ upon convergence are obtained substituting P in eq. (17). Diagonal values of matrix P_{filt} correspond to the variances of the displacement errors $\sigma^2_{\Delta d_i}$ of the WZ frames into the GOP. Intuitively, each $\sigma^2_{\Delta d_i}$ value represents the displacement error between the true motion and the estimated motion for the ith frame, which is needed to compute equation (4). Then, the average rate can be computed according to equation (1).

4. EXPERIMENTAL RESULTS

In order to run the simulations with the proposed model, we need to obtain realistic values of ρ and σ_d^2 for some test sequences. We performed motion estimation with $1/4$ pixel accuracy and we obtained the parameters of the AR(1) model (7) that best fits the estimated motion vectors along the motion trajectories.

Figure 2a-c depicts the rate-distortion curves obtained according to equation (1), indicating the estimated parameters ρ and σ_d^2 of the AR(1) model for the test sequences. The curves are calculated according to the following steps:

1. Set the GOP size N, the motion estimation accuracy σ_w^2, the state-space parameters (σ_d^2, ρ) and the spatial spectral density function ($\omega_0 = \pi/45$ as suggested in [2].)
2. Obtain the displacement error variances $\sigma^2_{\Delta d_i}$ by computing the trace of the matrix in equation (17).
3. For each value of θ:
   - Compute $R^I(\theta)$, $D^I(\theta)$ for the first frame of the GOP (intra-coded key frame) using equations (3) and (2).
   - For each Wyner-Ziv frame $i = 2, \ldots, N$
     - Obtain the power spectral density of the prediction error $\tilde{\phi}_{ee_i}(\Lambda)$, given $\sigma^2_{\Delta d_i}$ and $\phi_{ss}(\Lambda)$.
     - Compute the rate-distortion point corresponding to θ using equations (6) and (5)
   - Compute the average rate-distortion point according to equation (1).

Figure 2a-c shows that, based on the proposed model, motion-compensated prediction at the encoder outperforms motion-compensated interpolation at the decoder for the studied sequences. In fact, the lack of the original frame when generating the side information introduces a coding efficiency loss.

In addition, the optimal GOP size might depend on the target distortion. At high bit-rates, shorter GOP sizes are usually preferred. In fact, high frequencies are preserved, and accurate displacement estimation is needed to reduce the energy of the prediction error. In fact, as GOP size increases the displacement error variance also increases, thus impairing the accuracy of displacement estimation. Nevertheless, at low bit-rates, quantization filters out high frequencies, therefore a higher displacement error variance can be tolerated. This implies that the GOP size can be increased to reduce the number of intra-coded key frames.

We can conclude that the optimal GOP size depends on the underlying motion statistics. For sequences characterized by simple and temporally coherent motion like *Salesman*, the proposed model suggests that the optimal GOP size is between 4 and 8 frames. As the motion complexity increases (σ_d^2 increases), and the motion temporal coherence vanishes (ρ decreases), the optimal GOP size can be as little as 2 frames (see Figure 2c). For sequences characterized by very complex motion, it can also happen that pure intra-frame coding (i.e. GOP size equal to 1) outperforms Wyner-Ziv coding.

In order to validate the proposed model, we obtained the rate-distortion functions for the first 64 frames of some test sequences (*Salesman*, *Mother* and *Foreman*) at QCIF resolution and 15fps (see Figure 2d-f). The INTRA and INTER curves refer respectively to H.263+ intra (I-I-I) and H.263+ inter (I-P-P, GOP size 32). For the other curves, we adopted the motion-compensated interpolation algorithm described in [8], where the minimum block size is set equal to 16×16.

In order to isolate the impact of the generation of the side information alone, we replaced Turbo coding with conventional DCT-based intra-frame entropy coding of the prediction residuals as in H.263+. Therefore, we are providing results for a pseudo DVC-based coding architecture, where other design parameters that might affect the coding efficiency (i.e. correlation channel estimation, stopping criteria for Turbo decoding, encoder side rate-control) are explicitly singled out. In other words, the results provided can be interpreted as upper bounds that can be achieved if channel coding tools match the same performance of conventional entropy coding, when the formers are used for source coding.

By comparing the top and the bottom rows of Figure 2 we notice that coding efficiency of motion-compensated interpolation at the decoder falls in-between intra and inter-frame coding. Sometimes, it also falls below the intra-frame coding curve for long GOP sizes and sequences characterized by complex motion. Nevertheless, the coding efficiency of inter-frame coding is never achieved, suggesting that the lack of the current frame when generating the side information introduces a significant coding efficiency loss with

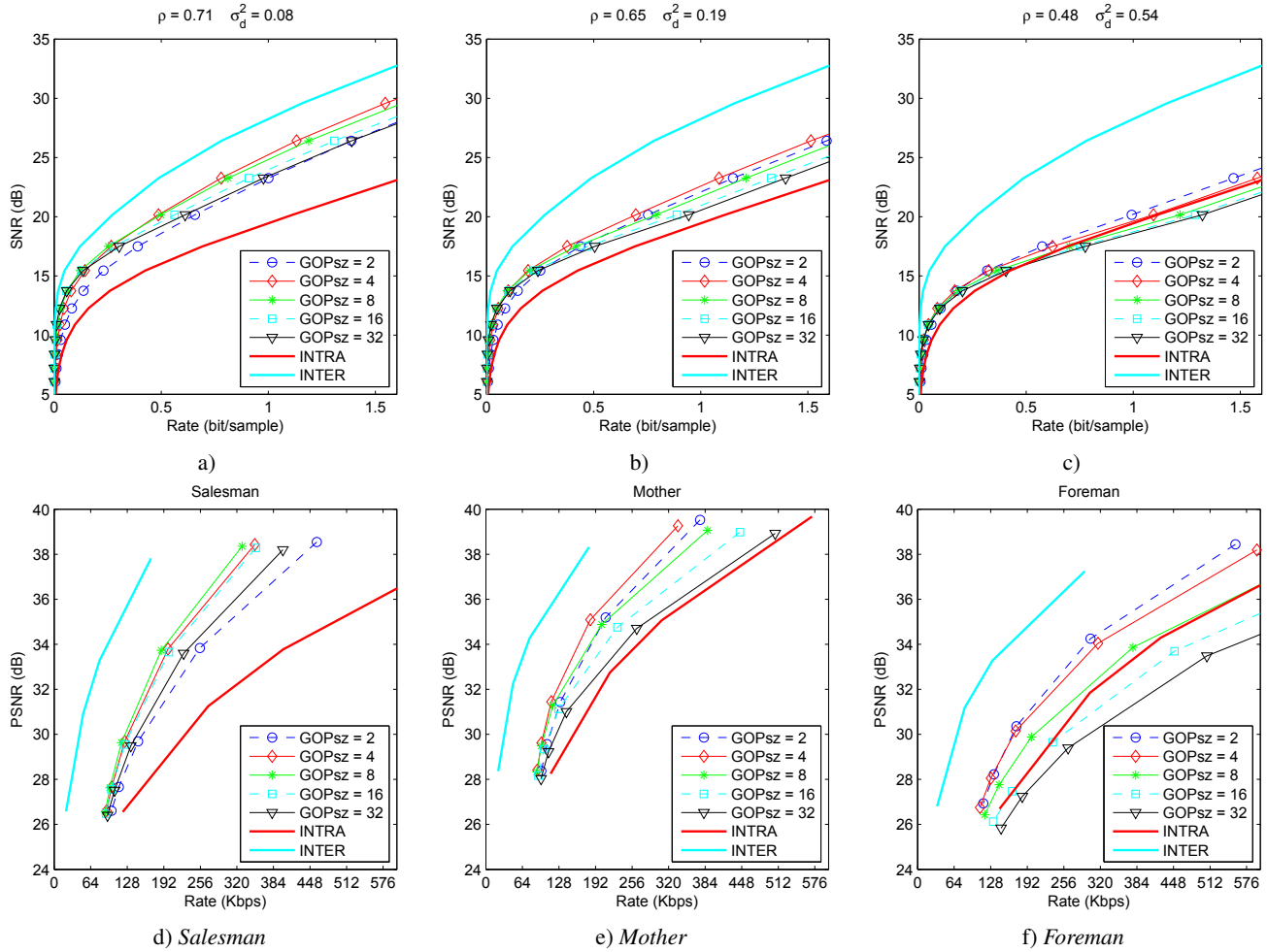

Fig. 2. (a-c) Rate-distortion curves obtained with the proposed model. Each plot indicates the values of ρ and σ_d^2 used to obtain the curves, estimated for the test sequences *Salesman*, *Mother* and *Foreman*. (d-f) Rate-distortion curves obtained for the test sequences.

respect to conventional motion-compensated predictive coding. In addition, the proposed model provides a quite accurate indication of the optimal GOP size for each of the tested sequences ($4-8$ for *Salesman*, 4 for *Mother* and 2 for *Foreman*). The difference between different GOP sizes can be better appreciated at high bit-rates, as suggested by the proposed model.

5. CONCLUSIONS

In this paper we propose a model that describes the rate-distortion characteristic of DVC-based coding schemes that perform motion-compensated interpolation at the decoder. Both the model simulations and the experiments on real video sequences show that the coding efficiency of inter-frame coding is not achieved. In addition, the optimal GOP size depends on the sequence motion complexity, typically ranging between 2 and 8 for the tested sequences.

6. REFERENCES

[1] Bernd Girod, Anne Aaron, Shantanu Rane, and David Rebollo Monedero, "Distributed video coding," *Proceedings of the IEEE*, vol. 93, pp. 71–83, January 2005.

[2] Bernd Girod, "The efficiency of motion-compensated prediction for hybrid coding of video sequences," *IEEE Journal on Selected Areas in Communications*, vol. 7, pp. 1140–1154, August 1987.

[3] Marco Tagliasacchi, Stefano Tubaro, and Augusto Sarti, "On the modeling of motion in Wyner-Ziv video coding," in *Proceedings of the International Conference on Image Processing*, Atlanta, GA, October 2006.

[4] Zhen Li and Edward J. Delp, "Wyner-Ziv video side estimator: Conventional motion search methods revisited," in *Proceedings of the International Conference on Image Processing*, Genova, Italy, September 2005, pp. 825 – 828.

[5] Anne Aaron and Bernd Girod, "Compression with side information using turbo codes," in *Proceedings of the IEEE Data Compression Conference*, Snowbird, UT, April 2002.

[6] João Ascenso, Catarina Brites, and Fernando Pereira, "Interpolation with spatial motion smoothing for pixel domain distributed video coding," in *EURASIP Conference on Speech and Image Processing, Multimedia Communications and Services*, Slovak Republic, July 2005.

[7] Toby Berger, *Rate Distortion Theory*, Prentice Hall, 1971.

[8] Luca Piccarreta, Augusto Sarti, and Stefano Tubaro, "An efficient video rendering system for real-time adaptive playout based phyical motion field estimation," in *European Signal Processing Conference*, Antalya, Turkey, September 2005.