

Rate-Distortion Analysis of Motion-Compensated Interpolation at the Decoder in Distributed Video Coding

Marco Tagliasacchi, *Member, IEEE*, Laura Frigerio, and Stefano Tubaro, *Member, IEEE*

Abstract—This letter 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. Therefore, the true motion field necessary for this operation is not directly available, and the motion vectors must be estimated at the decoder side, thus introducing displacement estimation errors. The accuracy of the motion-compensated interpolation at the decoder depends on several factors: 1) the overall motion complexity; 2) the temporal coherence of the motion field; and 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. In addition, the model provides a good estimate of the group of pictures size that optimizes the coding efficiency. Experimental results on real video sequences validate the results of the proposed model.

Index Terms—Distributed video coding, motion analysis, video coding.

I. INTRODUCTION

DISTRIBUTED video coding (DVC) is a recent video coding paradigm whose main idea is to perform intra-frame encoding and inter-frame decoding. The computational burden due to motion estimation is thus shifted to the decoder side. DVC is based on the principles of distributed source coding stated by the Slepian–Wolf [1] theorem for the lossless case and later extended by the Wyner–Ziv [2] theorem to the lossy scenario. 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 [3].

The goal of this letter 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 informa-

tion at the decoder by performing motion-compensated interpolation, starting from two intra-coded key frames [3]. 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 [4]. Armed with the proposed model, we investigate the tradeoff between motion-compensated interpolation accuracy and GOP size, in order to find the optimal group of pictures (GOP) size for a target distortion.

This letter extends our previous work in [5] 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 [6], where the model explicitly addresses only the case of motion-extrapolation.

II. 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 [3]. 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 performing motion estimation using $s(t)$ as current frame and the previously encoded frame $s'(t-1)$ as reference frame (s' is the quantized version of s). An $I - P - P - \dots - I$ GOP structure is assumed.
- *Motion-compensated interpolation at the decoder:* The decoder performs motion-compensated interpolation using lossy coded key frames $s'(\tau_1)$ and $s'(\tau_2)$ only ($\tau_1 < t < \tau_2$) [7], [8] to generate $\hat{s}(t) = \hat{s}_{WZ}(t)$. An $I - WZ - WZ - \dots - I$ GOP is adopted. 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

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

where $R^I(D)$ is the contribution of the intra-coded frame and $R_i^{\{P,WZ\}}(D)$ that of the i th inter-coded frame (for the case of

Manuscript received December 8, 2006; revised January 22, 2007. This work was supported in part by the Italian Ministry of Education and Research (MIUR) under the project PRIN-“Robust video coding techniques based on distributed source coding” and by the European Union (EU) under Visnet II Network of Excellence. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. James E. Fowler.

The authors are with Dipartimento di Elettronica e Informazione, Politecnico di Milano, 32 20133-Milano, Italy (e-mail: marco.tagliasacchi@polimi.it; laura.frigerio@polimi.it; stefano.tubaro@polimi.it).

Digital Object Identifier 10.1109/LSP.2007.896187

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 [9]:

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

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

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. We notice that when $\theta < \phi_{ss}(\Lambda) \forall \Lambda$, $D^I(\theta) = \theta$. Therefore, θ is proportional to the amount of distortion introduced by quantization.

In the following, we derive the rate-distortion curves $R^{\{P, WZ\}}(D)$ adopting the framework introduced in [4]. 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 [4]

$$\phi_{ee}(\Lambda) \approx \begin{cases} \phi_{ss}(\Lambda), & \text{if } \phi_{ss}(\Lambda) < \theta \\ \max\{\tilde{\phi}_{ee}(\Lambda), \theta\}, & \text{otherwise} \end{cases} \quad (4)$$

$$\tilde{\phi}_{ee}(\Lambda) = 2\phi_{ss}(\Lambda) \left(1 - e^{-\frac{1}{2}(\omega_x \sigma_{\Delta d_x}^2 + \omega_y \sigma_{\Delta d_y}^2)}\right) + \theta \quad (5)$$

where $\sigma_{\Delta d_c}^2$ 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 [4], an approximation of the rate-distortion function is given by

$$\begin{aligned} D^{\{P, WZ\}}(\theta) &= E[(e' - e)^2] \\ &= \frac{1}{4\pi^2} \iint_{\Lambda} \min[\theta, \phi_{ss}(\Lambda)] d\Lambda \end{aligned} \quad (6)$$

$$\begin{aligned} R^{\{P, WZ\}}(\theta) &= \frac{1}{8\pi^2} \iint_{\Lambda: (\phi_{ss}(\Lambda) > \theta \text{ and } \tilde{\phi}_{ee}(\Lambda) > \theta)} \log_2 \frac{\tilde{\phi}_{ee}(\Lambda)}{\theta} d\Lambda. \end{aligned} \quad (7)$$

We can observe that, in order to compute (1), we need to characterize the values of the displacement error variances $\sigma_{\Delta d_x}^2$ and $\sigma_{\Delta d_y}^2$ for each frame in the GOP. Assuming isotropic displacement errors, we can state that, on average, $\sigma_{\Delta d_x}^2 = \sigma_{\Delta d_y}^2 = \sigma_{\Delta d}^2$. Therefore, we will drop the coordinate index x, y in the rest of this letter. 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 ($M = 1, 1/2, 1/4, \dots$ pixels). Therefore, we can write $\sigma_{\Delta d}^2 = M^2/12$ for any frame in the GOP as indicated in [4].
- *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 $\sigma_{\Delta d_i}^2$ for the i th frame we propose a model based on Kalman filtering, detailed in the following section.

III. STATE-SPACE 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:

$$d(t) = \rho d(t-1) + z(t) \quad (8)$$

where $d(t)$ is the true displacement that the frame $s(t)$ is subject to, ρ is the temporal correlation coefficient and $z(t)$ is a zero-mean white noise, having variance σ_z^2 ($z(t) = WN(0, \sigma_z^2)$). The variance of $d(t)$ can be computed as $\sigma_d^2 = \sigma_z^2 / (1 - \rho^2)$. In order to gain an insight, we can interpret σ_d^2 as an indication of the motion complexity; and ρ as a measure of the temporal coherence of the motion field, for a given value of σ_d^2 . 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 $t, t-1, \dots, t-N+1$ (i.e., $\hat{d}(t), \hat{d}(t-1), \dots, \hat{d}(t-N+1)$), when only the motion $o(t)$ between two key frames is observed

$$o(t) = d(t) + d(t-1) + d(t-2) + \dots + d(t-N+1) + w(t) \quad (9)$$

where $w(t)$ is a white noise $WN(0, \sigma_w^2)$ that takes into account the finite accuracy of displacements ($\sigma_w^2 = M^2/12$), as already explained for *P* frames in the previous section. Equation (9) describes the fact that the true motion between two successive key frames can be expressed as the sum of displacements of the frames in between.

The state-space model described by (8) and (9), 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 Fig. 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 (8) and denoting $d_i(\tau) = d(\tau-1+k_i)$ and $z_i(\tau) = z(\tau-1+k_i)$, we obtain the subsequent model

$$\begin{aligned} d_1(\tau) &= \rho d(\tau-1) + z_1(\tau) \\ d_2(\tau) &= \rho^2 d(\tau-1) + \rho z_1(\tau) + z_2(\tau) \\ d(\tau) &= \rho^3 d(\tau-1) + \rho^2 z_1(\tau) + \rho z_2(\tau) + z(\tau) \\ o(\tau) &= d_1(\tau) + d_2(\tau) + d(\tau) + w(\tau) \end{aligned} \quad (10)$$

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

$$\mathbf{d}(\tau) = \mathbf{F} \mathbf{d}(\tau-1) + \mathbf{v}_1(\tau) \quad (11)$$

$$o(\tau) = \mathbf{H} \mathbf{d}(\tau) + v_2(\tau) \quad (12)$$

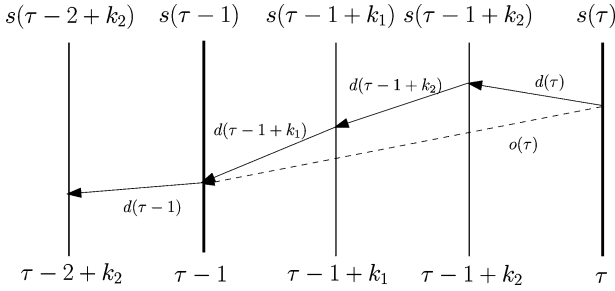


Fig. 1. Motion-compensated interpolation with time step τ referred to the evolution of the intra-coded key frames.

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 obtaining the following matrices:

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

$$H_{(1 \times N)} = (1 \quad 1 \quad \cdots \quad 1)$$

$$V_{1(N \times N)} = \sigma_z^2 \begin{pmatrix} 1 & \rho & \cdots & \rho^{N-1} \\ \rho & \rho^2 + 1 & \cdots & \rho^N + \rho^{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{N-1} & \rho^N + \rho^{N-2} & \cdots & \sum_{i=1}^N \rho^{2(N-i)} \end{pmatrix}$$

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

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 $\sigma_{\Delta d_i}^2$ of the i th 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 Riccati Differential Equation (RDE)

$$P(\tau+1) = FP(\tau)F^T + V_1 - K(\tau)(HP(\tau)H^T + V_2)K^T \quad (13)$$

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

$$\begin{aligned} E[\Delta \mathbf{d}(\tau|\tau)\Delta \mathbf{d}^T(\tau|\tau)] \\ &= P_{filt}(\tau) \\ &= P(\tau) - P(\tau)H^T[(HP(\tau)H^T) + V_2]^{-1}HP(\tau). \end{aligned} \quad (14)$$

In (13), upon convergence, $P(\tau+1) = P(\tau) = P$. Substituting P into (13), we obtain the Algebraic Riccati Equation (ARE) and we solve by P . Values of the matrix $P_{filt}(\tau)$ upon convergence are obtained substituting P in (14). Diagonal values of matrix P_{filt} correspond to the variances of the displacement errors $\sigma_{\Delta d_i}^2$ of the WZ frames into the GOP. Intuitively, each $\sigma_{\Delta d_i}^2$

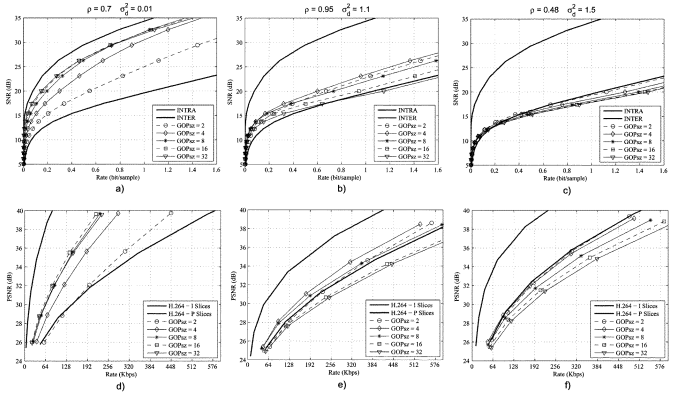


Fig. 2. (a)–(c) Rate-distortion curves obtained with the proposed model. Each plot indicates the values of ρ and σ_d^2 used to obtain the curves. (d)–(f) Rate-distortion curves obtained for real test sequences *Salesman*, *Coastguard*, and *Foreman*.

value represents the displacement error between the true motion and the estimated motion for the i th frame, which is needed to compute (4). Then, the average rate can be computed according to (1).

IV. EXPERIMENTAL RESULTS

In order to obtain realistic values of ρ and σ_d^2 to be used in the model simulations, we computed them for some test sequences. We performed motion estimation with 1/4 pixel accuracy and we obtained the parameters of the AR(1) model (8) that best fits the estimated motion vectors along the motion trajectories.

Fig. 2(a)–(c) depicts the rate-distortion curves obtained according to (1), indicating the estimated parameters ρ and σ_d^2 of the AR(1) model (8) for the test sequences. The curves are calculated according to the following steps.

- 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 (we use the isotropic PSD $\phi_{ss}(\Lambda)$ suggested in [4].)
- Obtain the displacement error variances $\sigma_{\Delta d_i}^2$ by computing the trace of the matrix in (14).
- For each value of θ :
 - Compute $R^I(\theta)$, $D^I(\theta)$ for the first frame of the GOP (intra-coded key frame) using (3) and (2).
 - For each Wyner-Ziv frame $i = 2, \dots, N$.
 - Obtain the power spectral density of the prediction error $\check{\phi}_{e_i}(\Lambda)$, given $\sigma_{\Delta d_i}^2$ and $\phi_{ss}(\Lambda)$.
 - Compute the rate-distortion point corresponding to θ using (6) and (7).
 - Compute the average rate-distortion point according to (1).

The INTER curve is obtained with (6) and (7), where $\sigma_{\Delta d}^2 = M^2/12$ ($M = 4$) regardless of the frame index, as explained in Section II.

Fig. 2(a)–(c) shows that motion-compensated interpolation is unable to achieve the coding efficiency of conventional motion-compensated prediction at the encoder for the studied sequences. Therefore, 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 pre-

diction error. As the 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. 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 can be as large as 16 to 32 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 1 to 2 frames [see Fig. 2(c)]. Therefore, for sequences characterized by complex motion like *Foreman*, 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 some test sequences (*Salesman*—192 frames, *Coastguard*—128 frames, and *Foreman*—192 frames) at QCIF resolution and 15 fps. Results are provided for H.264/AVC, using either I-slices (I-I-I) or P-slices (I-P-P, GOP size 32). For the other curves, we adopted the motion-compensated interpolation algorithm described in [10], where the minimum block size is set equal to 4×4 .

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.264/AVC. 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 Fig. 2(a)–(c) with Fig. 2(d)–(f), 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 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 (16 for *Salesman*, 4 for *Coastguard* and 1, i.e., pure intra-frame coding, for *Foreman*). The difference between different GOP sizes can

be better appreciated at high bit rates, as suggested by the proposed model.

V. CONCLUSIONS

In this letter, 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 motion-compensated predictive inter-frame coding is not achieved. This is due to the fact that the lack of the original frame introduces a displacement estimation error, which is not only limited by the finite accuracy of motion vectors. We showed that the displacement error depends on the GOP size and on the intrinsic motion characteristics of the sequence. In addition, the proposed model provides a good estimate of the optimal GOP size. We argue that the combined use of motion-compensated interpolation with other coding tools (frame hashes, motion refinement, etc.) might partially bridge the gap observed in this letter. This is a subject of current investigations.

REFERENCES

- [1] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. IT-19, pp. 471–480, Jul. 1973.
- [2] A. D. Wyner and J. Ziv, "The rate distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.
- [3] B. Girod, A. Aaron, S. Rane, and D. R. Monedero, "Distributed video coding," *Proc. IEEE*, vol. 93, no. 1, pp. 71–83, Jan. 2005.
- [4] B. Girod, "The efficiency of motion-compensated prediction for hybrid coding of video sequences," *IEEE J. Sel. Areas Commun.*, vol. SAC-5, no. 7, pp. 1140–1154, Aug. 1987.
- [5] M. Tagliasacchi, S. Tubaro, and A. Sarti, "On the modeling of motion in Wyner-Ziv video coding," in *Proc. Int. Conf. Image Processing*, Atlanta, GA, Oct. 2006.
- [6] Z. Li and E. J. Delp, "Rate distortion analysis of motion side estimation in Wyner-Ziv video coding," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 98–113, Jan. 2007.
- [7] A. Aaron and B. Girod, "Compression with side information using turbo codes," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Apr. 2002.
- [8] J. Ascenso, C. Brites, and F. Pereira, "Interpolation with spatial motion smoothing for pixel domain distributed video coding," in *Proc. EURASIP Conf. Speech and Image Processing, Multimedia Communications and Services*, Slovak Republic, Jul. 2005.
- [9] T. Berger, *Rate Distortion Theory*. Upper Saddle River, NJ: Prentice-Hall, 1971.
- [10] L. Piccarreta, A. Sarti, and S. Tubaro, "An efficient video rendering system for real-time adaptive playout based physical motion field estimation," in *Proc. Eur. Signal Processing Conf.*, Antalya, Turkey, Sep. 2005.