

HASH-BASED MOTION MODELING IN WYNER-ZIV VIDEO CODING

Marco Tagliasacchi, Stefano Tubaro

Dipartimento di Elettronica e Informazione
Politecnico di Milano
P.zza Leonardo da Vinci, 32 20133 - Milano, Italy

ABSTRACT

Generally, Distributed video coding (DVC) schemes perform motion estimation at the decoder side, without the current frame being available. In order to generate the side-information reliably, one solution consists in allocating a limited bit budget to send a hash of the current frame. At the decoder, this auxiliary hash is used to perform motion estimation. This paper studies the accuracy of hash-based motion estimation and compares it to conventional encoder-side motion estimation. We show that, at low rates, the very limited bit-budget of the hash does not ensure a reliable motion estimation, while at medium to high rates the motion accuracy is comparable with the finite precision used to represent motion vectors. Then, we derive the rate-distortion characteristic, which combines the cost of encoding the hash and the prediction residuals after decoder-side motion compensation. We show that, at high rates, hash-based motion modeling can virtually achieve the same coding efficiency as motion-compensated predictive coding. Instead, at medium-to-low rates we observe a significant coding loss. Experimental results on real video sequences validate the results of the proposed model.

Index Terms— Video coding, motion analysis, rate distortion theory

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. In fact, the computational burden due to motion estimation is shifted to the decoder side. DVC is based on the principles of distributed source coding stated by the Slepian-Wolf[1] and Wyner-Ziv[2] theorems.

Results obtained on test video sequences reveal that DVC coding schemes generally improve the coding efficiency with respect to intra-frame coding, but they are unable to achieve the gains of conventional motion-compensated predictive codecs [3][4]. The coding efficiency gap can be attributed to different reasons: sub-optimality of channel coding tools (Turbo, LDPC); inaccuracies in the correlation noise modeling between the source to be decoded and the side information; mismatch between Wyner-Ziv theorem hypothesis and the video coding scenario (non-Gaussian noise, finite block lengths); sub-optimality of motion modeling when performed at the decoder.

In this paper we focus our attention on the latter element only. In conventional motion-compensated predictive codecs, a large part of the coding efficiency gain achieved in past twenty years is due to more accurate motion models. Such models are obtained at the encoder by comparing the current frame to be encoded with one or more reference frames. In the case of DVC-based solutions, the decoder computes the motion model without the current frame itself being available. Therefore, the side information Y available at the

encoder side, i.e. the best motion compensated prediction of the current frame, is not available at the decoder. In general, a worse version of Y , say \hat{Y} ($E[(X - Y)^2] \leq E[(X - \hat{Y})^2]$), can be generated. This results in a coding efficiency loss.

In the literature, a number of practical solutions have been proposed in order to improve the quality of the side information generated at the decoder, in such a way to make $\hat{Y} \sim Y$. Refined motion interpolation/extrapolation schemes [5][6], motion estimation based on auxiliary hash functions [7][8][4], cyclic redundancy checksums (CRC's) [9] have been thoroughly studied. This paper studies the motion model accuracy computed by means of hash functions, and relates this to the rate-distortion performance of the overall coding scheme.

The rest of the paper is organized as follows. Section 2 introduces the motion models currently used in state-of-the-art Wyner-Ziv coding schemes. Section 3 briefly summarizes the main results in [10], which will be used in the rest of the paper. The motion model accuracy is studied in Section 4 and it is related to the overall rate-distortion performance of the coding scheme in Section 5. Experimental results validating the proposed model are illustrated in Section 6.

2. HASH-BASED MOTION MODELS IN WYNER-ZIV VIDEO CODING

Let us denote the frame to be encoded as $s(t)$ and its motion compensated predictor (side information) as $\hat{s}(t)$. Figure 1 illustrates the generic input of the side information generator module. Depending on the available information, we can distinguish two cases:

- a) *motion estimation at the encoder*: $\hat{s}_a(t)$ is generated at the encoder side. The current frame $s(t)$ is available together with the previously encoded frame $s(t - 1)$ and its reconstructed version $s'(t - 1)$ ¹;
- b) *hash based motion estimation at the decoder*: $\hat{s}_b(t)$ is generated at the decoder side. The current frame is not available, but the encoder sends a quantized hash function of $s(t)$, $s'_h(t)$. The hash may consist of low frequency DCT coefficients [7], or high frequency DCT coefficients [8]. The decoder performs motion estimation comparing $s'_h(t)$ with the previously decoded frame $s'(t - 1)$, then uses the estimated motion model to perform motion compensation computing $\hat{s}_b(t)$.

Let us denote the residual frame after motion-compensated prediction as $e(t) = s(t) - \hat{s}(t)$. Define the power spectral density

¹ x' denotes the quantized version of x reconstructed at the decoder.

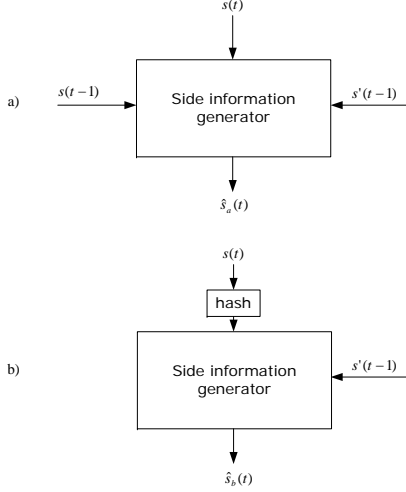


Fig. 1. a) Motion estimation at the encoder. b) Hash based motion estimation at the decoder.

(PSD) of $e(t)$ as $\phi_{ee}(\Lambda)$ ($\Lambda = (\omega_x, \omega_y)$). In the following section we relate $\phi_{ee}(\Lambda)$ to the motion model accuracy and to the signal power spectral density $\phi_{ss}(\Lambda)$, thus deriving the rate-distortion function according to [10].

3. INFORMATION THEORETIC BACKGROUND

Let us assume that the signal to be encoded s is a stationary, jointly Gaussian, zero-mean signal with power spectral density $\phi_{ss}(\Lambda)$. The reconstructed version at the decoder is denoted by s' . Then, for a given mean-square error [11]

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

the minimum transmission rate that can be achieved is

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

where $\theta > 0$ is a real-valued parameter that allows to sweep the rate-distortion curve. When $\theta < \phi_{ss}(\Lambda) \forall \Lambda$, $D = \theta$.

In the inter-frame coding case, the prediction error signal e is encoded instead of s . In [10], an approximation of the rate-distortion function is given by

$$D(\theta) = E[(s' - s)^2] = E[(e' - e)^2] = \frac{1}{4\pi^2} \iint_{\Lambda} \min[\theta, \phi_{ss}(\Lambda)] d\Lambda \quad (3)$$

$$R(\theta) = \frac{1}{8\pi^2} \iint_{\Lambda: (\phi_{ss}(\Lambda) > \theta \text{ and } \phi_{ee}(\Lambda) > \theta)} \log_2 \frac{\phi_{ee}(\Lambda)}{\theta} d\Lambda \quad \text{bit} \quad (4)$$

4. MOTION MODEL ACCURACY

Let us consider a video signal that contains exclusively a constant, translatory displacement (d_x, d_y) , and neglect any other effects like rotation, zoom, occlusions, illumination changes, etc. The expression of $\phi_{ee}(\Lambda)$ when no spatial prediction is performed is given by

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

where the random error between the actual and estimated displacement is defined by

$$\begin{bmatrix} \Delta d_x \\ \Delta d_y \end{bmatrix} = \begin{bmatrix} d_x \\ d_y \end{bmatrix} - \begin{bmatrix} \hat{d}_x \\ \hat{d}_y \end{bmatrix} \quad (6)$$

and the probability density function (p.d.f.) $p_{\Delta d_x, \Delta d_y}(\Delta d_x, \Delta d_y)$ is assumed to Gaussian with covariance matrix $\text{diag}[\sigma_{\Delta d_x}^2, \sigma_{\Delta d_y}^2]$. In [10] it is observed that the rate-distortion characteristic is not very sensitive to the shape of the displacement error p.d.f., but on its variance.

When motion estimation is performed at the encoder, we can assume that the motion model accuracy is solely limited by the finite precision used to represent motion vectors. Conventional motion-compensated predictive codecs use a fractional-pel precision Δ ($\Delta = 1, 1/2, 1/4, \dots$) for both the horizontal and vertical components. The displacement error variance is given by $\sigma_x^2 = \sigma_y^2 = \Delta^2/12$.

In [7] motion estimation at the decoder has access to auxiliary information under the form of a hash function of the frame to be decoded. The underlying idea is that a partial description of the current frame may help in obtaining a more accurate estimate of the motion model. In [7] the hash consists in high-frequency coefficients. These hash bits serve to relay the motion information to the decoder without actually estimating the motion at the encoder.

In the proposed model, we denote the hash as

$$s_h(t) = (s * h)(t) \quad (7)$$

where $h(x, y)$ is the impulse response of an ideal high-pass. The corresponding frequency response is denoted by $H(\Lambda)$.

$$H(\Lambda) = H(\omega_x, \omega_y) = H(\omega_x)H(\omega_y) \quad (8)$$

$$H(\omega) = \begin{cases} 1 & \omega > F\pi \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where F is a real number in $[0, 1]$. We note that this expression of $H(\Lambda)$ only approximates the fact that some of the DCT coefficients are discarded.

The filtered version of the original signal, $s_h(t)$, is encoded and transmitted to the decoder as $s'_h(t)$. If we assume optimum intra-frame encoding of $s_h(t)$, according to equations (1) and (2), $s'_h(t)$ can be expressed as the output of the optimum forward channel shown in Figure 2.

Consecutive frames differ only by a constant, unknown displacement (d_x, d_y) . Thus, we can write

$$s(x, y, t) = s(x - d_x, y - d_y, t - 1); \quad (10)$$

and,

$$\begin{aligned} s'_h(x, y, t) &= (g * s_h)(x, y, t) + n_0(x, y, t) \\ &= w(x, y, t) + n_0(x, y, t) \end{aligned} \quad (11)$$

$$\begin{aligned} s'_h(x, y, t - 1) &= (g * s_h)(x + d_x, y + d_y, t) + n_1(x, y, t) \\ &= w(x + d_x, y + d_y, t) + n_1(x, y, t) \end{aligned} \quad (12)$$

where n_0 and n_1 are quantization noise terms characterized by a power spectral density

$$\phi_{nn}(\Lambda) = \phi_{n_0 n_0}(\Lambda) = \phi_{n_1 n_1}(\Lambda) = \max \left[0, \theta \left(1 - \frac{\theta}{\phi_{s_h s_h}} \right) \right] \quad (13)$$

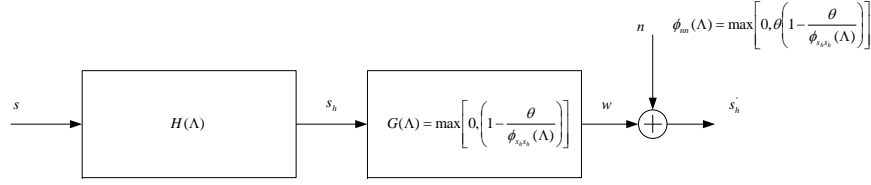


Fig. 2. Optimum forward channel corresponding to the rate-distortion function (1), (2)

and $g(x, y)$ is the impulse response corresponding to the transfer function $G(\Lambda)$ of the optimal forward channel (see Figure 2).

The problem becomes that of estimating the displacement (dx, dy) given the noisy observations $s'_h(x, y, t)$ and $s'_h(x, y, t - 1)$. The solution to this problem has been studied in the field of time delay estimation (TDE). If we assume that displacements are uncorrelated, the problem can be expressed in 1-d. The variance of the displacement error is [12]

$$\sigma_{\Delta d}^2 = \frac{2 \int_{-\pi}^{+\pi} \omega^2 \phi_{nn}(\omega) \phi_{ww}(\omega) d\omega + \int_{-\pi}^{+\pi} \omega^2 (\phi_{nn}(\omega))^2 d\omega}{N_b \left[\int_{-\pi}^{+\pi} \omega^2 \phi_{ww}(\omega) d\omega \right]^2} \quad (14)$$

where $\phi_{ww}(\omega) = |H(\omega)|^2 |G(\omega)|^2 \phi_{ss}(\omega)$ and N_b is the number of samples used to estimate the displacement. In our case N_b is the block size used for motion estimation. In addition, motion vectors are represented with a finite accuracy, thus

$$\sigma_{\Delta d}^2 = \max\{\Delta^2/12, \sigma_{\Delta d}^2\} \quad (15)$$

At high rates, $\theta \rightarrow 0$ (thus, according to (1), $D \rightarrow 0$), and the noise power spectrum is white $\phi_{nn}(\omega) = \theta$ (see equation (13)). In this scenario, the displacement error variance depends on

- the signal-to-noise ratio, which is driven by the quantization noise;
- the effective bandwidth β^2 of the signal w , defined as

$$\beta^2 = \int_{-\pi}^{+\pi} \omega^2 \phi_{ww}(\omega) d\omega, \quad (16)$$

which depends on the spatial spectral properties of the encoded sequence;

- the number of samples N_b , which in practice is limited by the fact that the motion is only locally constant.

In the low-to-medium rate range, quantization affects the accuracy of motion estimation in two ways:

- decreases the signal-to-noise ratio ρ , as already noted before;
- modifies the power spectral density of $\phi_{ww}(\omega)$, reducing the range of frequencies over which $\phi_{ww}(\omega) > 0$. Specifically, the filter $G(\omega)$ sets to zero frequencies where $\phi_{s_h s_h}(\omega) < \theta$ and attenuates frequencies where $\phi_{s_h s_h}(\omega) \simeq \theta$. The effective bandwidth is therefore decreased, since part of the high frequency content is lost in the quantization process. This fact, in turns, increases the variance of the displacement error $\sigma_{\Delta d}^2$.

5. RATE-DISTORTION ANALYSIS

In [7], the encoder sends a hash that consists of high-frequency DCT coefficients. At the decoder, these DCT coefficients are used to build

a high-pass approximation of the frame to be decoded and the latter is used for motion estimation. The rationale behind this idea is that high-frequency DCT coefficients contain information about the image edges, thus turning motion estimation into the problem of edge-based image registration. Based on the model described in Section 4, a hash that consists of high frequencies has a larger effective bandwidth, at least at high rates, thus a smaller displacement error variance σ_{Δ}^2 .

With the model introduced in Section 4, the coding algorithm can be summarized as follows

- set θ to achieve the desired distortion level. This corresponds to choosing the quantization step size;
- let $H(\omega)$ be a high-pass filter with cut-off frequency $F\pi$;
- the encoder computes and encode $s_h(t)$. Optimal intra-frame encoding of $s_h(t)$ is performed according to equation (2) and (1);
- the decoder decodes an approximation of $s_h(t)$, $s'_h(t)$;
- the decoder performs motion estimation using $s'_h(t)$ as current frame and $s'_h(t - 1)$ as reference frame. The displacement error is given by (14);
- the encoder encodes the low-frequency part of $s(t)$. This is equivalent to performing inverse water-filling on the power spectral density

$$\phi_{ee}(\Lambda) = 2\phi_{ss}(\Lambda) |1 - H(\Lambda)|^2 (1 - e^{-\frac{1}{2}(\omega_x \sigma_{\Delta d_x}^2 + \omega_y \sigma_{\Delta d_y}^2)}) + \theta \quad (17)$$

according to equations (4) and (3), where $\sigma_{\Delta d_x}^2$ and $\sigma_{\Delta d_y}^2$ are given by (14).

6. EXPERIMENTAL RESULTS

In order to numerically evaluate the rate-distortion curves, the band-limited spatial power spectral density of the 2-d random sequence $\{s(x, y, \cdot)\}$ is set equal to:

$$\phi_{ss}(\omega_x, \omega_y) = \begin{cases} \frac{2\pi}{\omega_0^2} \cdot \left(1 + \frac{\omega_x^2 + \omega_y^2}{\omega_0^2}\right)^{-3/2} & \text{for } |\omega_x| \leq \pi \\ & \text{and } |\omega_y| \leq \pi \\ 0, & \text{otherwise} \end{cases} \quad (18)$$

where ω_0 is chosen equal to $\pi/45$, as suggested by [10].

Figure 3 shows the rate-distortion curves corresponding to hash-based motion modeling, for different values of F . Both intra-frame and inter-frame coding are shown for ease of comparison. At high rates, a coding efficiency gain is achieved by increasing F . By reducing the number of coefficients used to generate the hash ($F \rightarrow 1$), the coding efficiency approaches that of inter-frame coding. This means that at high signal-to-noise ratios, a small fraction of DCT

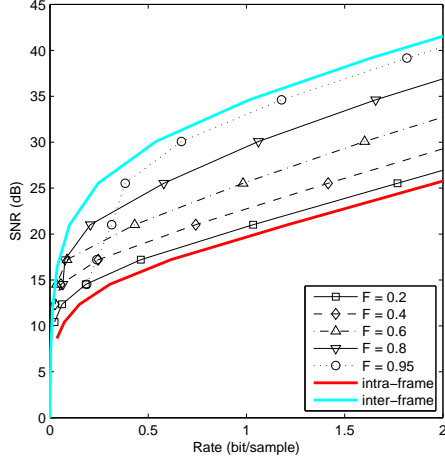


Fig. 3. Rate-distortion functions of hash-based motion modeling. Intra-frame vs. inter-frame hash encoding. a) Low-frequency hash. b) High-frequency hash.

coefficients is enough to achieve a motion model accuracy comparable with the finite precision of motion vectors.

Below approximately 0.5 bit / sample, quantization severely affects the motion model accuracy. We notice that, for a given value of F , there is a lower bound for the achievable bit-rate. Such a bound corresponds to the distortion level above which the power spectral density of the hash is reduced to nil.

Figure 4 shows the rate-distortion curves obtained by encoding the *Salesman* sequence, QCIF@15fps, GOP size equal to 8. We observe the same general behavior predicted by the proposed model. In fact, at high rates, the coding efficiency increases by reducing the number of DCT coefficients used to generate the hash. We notice two differences with respect to the proposed model:

- there is no lower bound on the achievable rate, even for values of $F \rightarrow 1$. In fact, in the actual codec implementation, the motion vector is set to zero whenever the hash consists of all zero coefficients. At very low rates, all the quantized coefficients of the hash tend to be zero, regardless of the value of F ;
- a coding efficiency gap still exists between inter-frame coding and hash based motion modeling. This can be attributed to the following reasons: 1) the actual codec generates the hash by retaining a subset of DCT coefficients. The reconstructed hash image at the decoder suffers from blocking artifacts, impairing the quality of motion estimation; 2) the actual motion is not purely translational, as assumed in the proposed model; 3) in the model, the sequence is assumed to be noise-free. In practice, acquisition noise affects the quality of motion estimation. The hash signal-to-noise ratio is lower than the original frame signal-to-noise ratio, due to the fact that most of the signal energy is concentrated at low frequencies.

7. CONCLUSIONS

The proposed model allows to describe the rate-distortion behavior of hash-based motion modeling. The goal is to capture the coding efficiency loss due to the fact that motion estimation is performed at the decoder side, relying on partial information on the current

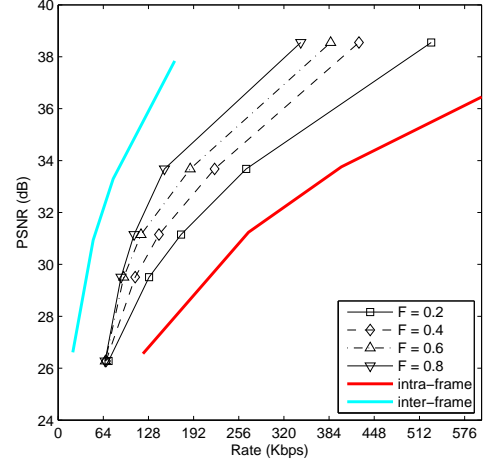


Fig. 4. Rate-distortion curves for the *Salesman* sequence.

frame. Our findings show that, at low-to-medium rates, quantization severely affects the motion estimation accuracy, causing a coding efficiency loss with respect to conventional inter-frame coding. The model described in this paper can be extended to address zero-motion inter-frame encoding of the hash information, as explained in [7]. Nevertheless, it turns out that the rate-distortion analysis depends also on the temporal correlation properties of the sequences. Only sequences characterized by little motion might benefit from this technique. In addition, we are studying the extension to low-frequency hash motion modeling [8] and Wyner-Ziv coding of residuals [4].

8. REFERENCES

- [1] David Slepian and Jack K. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, pp. 471–480, July 1973.
- [2] Aaron D. Wyner and Jacob Ziv, "The rate distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, pp. 1–10, January 1976.
- [3] Bernd Girod, Anne Aaron, Shantanu Rane, and David Rebollo Monedero, "Distributed video coding," *Proceedings of the IEEE*, vol. 93, pp. 71–83, January 2005.
- [4] Anne Aaron, David Varodayan, and Bernd Girod, "Wyner-Ziv residual coding of video," in *Picture Coding Symposium*, Beijing, April 2006.
- [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] Anne Aaron and Bernd Girod, "Wyner-ziv video coding with low-encoder complexity," in *Picture Coding Symposium*, San Francisco, CA, December 2004.
- [8] Anne Aaron and Bernd Girod, "Wyner-ziv video coding with hash-based motion compensation at the receiver," in *Proceedings of the International Conference on Image Processing*, Singapore, October 2004.
- [9] Rohit Puri and Kannan Ramchandran, "PRISM: A New Robust Video Coding Architecture based on Distributed Compression Principles," in *Allerton Conference on Communication, Control and Computing*, Urbana-Champaign, IL, October 2002.
- [10] 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.
- [11] T. Berger, *Rate Distortion Theory*, Prentice Hall, 1971.
- [12] Azizul H. Quazi, "An overview on the time delay estimate in active and passive systems for target localization," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, pp. 527–533, June 1981.