RATE-DISTORTION ANALYSIS AND BIT ALLOCATION STRATEGY FOR MOTION ESTIMATION AT THE DECODER USING MAXIMUM LIKELIHOOD TECHNIQUE IN DISTRIBUTED VIDEO CODING

Ivy H. Tseng and Antonio Ortega

Signal and Image Processing Institute, University of Southern California, Los Angeles, CA, 90089

ABSTRACT

Numerous approaches for distributed video coding have been recently proposed. One of main motivations for these techniques is the possibility of achieving complexity tradeoffs between the encoder and the decoder that may not be feasible in the context of conventional video coding. In our previous work, a Maximum Likelihood (ML) method for motion estimation at the decoder was proposed. It was shown that the ML method, designed based on the PRISM architecture, induces no additional rate cost, and is able to work with existing methods, e.g. those based on hash functions or CRC, to improve overall decoding PSNR. In this work, we present a rate-distortion analysis of our ML method. This analysis, given a correlation model for the video data, allows us to improve bit allocation at the encoder, i.e., the decision on the number of cosets to be used to represent various types of video information. We also signal "End of Block" (EOB) in coding to further exploit the energy compaction. Our experiments demonstrate significant improvements PSNR up to 1.5dB from RD optimized bit allocation.

Index Terms— Distributed video coding, Maximum likelihood, Rate-distortion analysis, Bit allocation

1. INTRODUCTION

Distributed Video Coding (DVC) techniques have been recently proposed based on distributed source coding (DSC) principles. One potential advantage of DVC is that it enables trade-offs between encoding and decoding complexity. In particular, motion estimation with reduced complexity can be performed at the encoder, while the decoder performs some form of search to exploit inter-frame correlations.

In many DSC applications, for each set of data to be encoded (e.g., one band in a hyperspectral image, as in [1]), there is a unique corresponding dataset available at the decoder that will be used as the side information (in [1], it is the previous band in a given encoding order). Instead, in the context of DVC, exploiting inter-frame correlation to reduce the rate is achieved by allowing the decoder to select, for each block within a frame, one of multiple blocks in the previous frame (corresponding to different motion displacements) as the side information for decoding. Thus, a major challenge in DVC is to develop techniques that allow the decoder to identify the best side information, given that the information sent by the encoder is ambiguous. This process can be seen as a form of motion estimation and compensation performed at the decoder.

Several practical methods have been proposed to enable motion estimation at the decoder. Aaron and Girod proposed sending a hash function containing auxiliary information of the original frame [2]. Puri and Ramchandran used CRC to validate the correctness of the decoded blocks [3]. Both methods require an explicit increase in transmission rate. In our previous work [4] a ML method was proposed. This method, designed for the PRISM architecture, exploits correlation statistics available at the decoder, in principle requires no auxiliary information, and can be used independently of or in combination with other motion estimation methods.

In the DVC literature, rate decisions are often made to meet a target decoding error probability. This requires taking into consideration the correlation between the data to be encoded and the side information available at the decoder, either at the frame level, as in systems using LDPC-based Slepian-Wolf encoders (e.g., [1, 2]), or at the block level (e.g., [3]). For example, for block-based systems, such as PRISM [3], the number of cosets used to encode each DCT coefficient is determined based on available statistics of the residual energy between the data and the side information.

However, in video coding problems, PSNR, rather than decoding error probability, is the conventional performance metric. In this paper, we present a rate-distortion analysis of the decoding performance using the ML method, with the mean squared error (MSE) of the reconstructed signal as the distortion measure. We express MSE as a function of the number of bits used to encode each DCT coefficient and of the given the statistical models. The analysis is then used to decide the number of bits to devote to each DCT coefficient in order to minimize MSE. This R-D analysis shows that both the characteristics of the correlation between data and side information, *and* the statistics of the data to be encoded affect MSE, and thus both should be considered in rate decision. We also introduce the EOB signal into our system to further exploit the energy compaction. The RD optimized bit allocation posts 0.7dB-1.5dB improvement in PSNR. The coding efficiency is also greatly improved with the introduction of the EOB method, particularly in the lower rate region, for up to 2.5dB gain in PSNR.

This paper is organized as follows. In Section 2 we briefly review the DVC system using the ML method proposed in [4]. In Section 3 we introduce our assumptions and the problem formulation. In Section 4 the rate-distortion analysis and the bit allocation strategy are presented. Experimental results are shown in Section 5 to validate our analysis.

2. SYSTEM OVERVIEW

In this section, we briefly review the ML method for motion estimation at the decoder and the corresponding video coding system proposed in our previous work [4]. This system, as shown in Figure 1, is an extension of a PRISM DVC, where a ML approach is used to identify blocks in the previous frame that should be used as side information at the decoder.

Let $X = \{x^d\}_{d=1,\cdots,D}$ and $Y_m = \{y_m^d\}_{d=1,\cdots,D}$ be D-dimensional vectors representing the DCT coefficients of the cur-

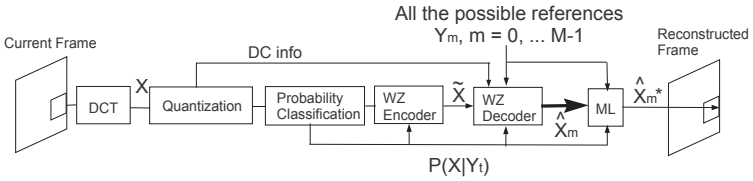

Fig. 1. The DVC system proposed in [4]. It is a PRISM-based transform domain distributed video coding architecture with ML decoding

rent block and the m-th reference block, respectively, where $m = 0, \cdots, M-1$ (i.e., M is the total number of reference blocks). The encoder has computed an estimate of the joint statistics of X and the true side information Y_t:

$$P(X|Y_t) = f_{X|Yt}(X, Y_t).$$

$P(X|Y_t)$ should be known at both encoder and decoder in order to enable efficient DSC. The encoder can perform motion estimation on some of the data to estimate the model. X is then Wyner-Ziv (WZ) coded as $\tilde{X}$ and transmitted along with $P(X|Y_t)$ and potentially some helper information (to be discussed further below). $\tilde{X}$ can be decoded using each of the references $Y_m, m = 0, \cdots, M-1$ as side information, with $\hat{X}_m$ representing the decoding result when Y_m is used. For each pair $(\hat{X}_m, Y_m)$ we compute the likelihood of $\hat{X}_m$ given Y_m based on $P(X|Y_t)$:

$$L(\hat{X}_m, Y_m) = f_{X|Yt}(\hat{X}_m, Y_m). \tag{1}$$

If

$$m^* = \arg\max_m L(\hat{X}_m, Y_m), \tag{2}$$

then Y_{m^*} is chosen as side information and the corresponding $\hat{X}_{m^*}$ will be the decoded block.

In our system two types of "helper information" are transmitted in addition to the WZ coded information. The DC value of each encoded block is sent directly (without WZ coding), since the DC values from neighboring blocks are usually highly correlated and thus can be coded very efficiently. We also employ an explicit EOB strategy, i.e., the location of the last non-zero coefficient in each block (when scanned in a zigzag fashion) is signaled to exploit the energy compaction properties in the transform domain. Thus, the zero coefficients at the end of each block are not encoded using WZ techniques. Note that both these types of intra information can be used to help in identifying the correct side information at the decoder.

3. PROBLEM STATEMENT

Our goal in this paper is to derive a rate-distortion model when our proposed ML method is used, for given statistical models and encoding options. We assume the following model for the joint statistics of X and Y_t:

$$X = Y_t + N_t \tag{3}$$

where N_t is a noise term and is independent of Y_t. Both $Y_t = \{y_t^d\}_{d=1,\cdots,D}$ and $N_t = \{n_t^d\}_{d=1,\cdots,D}$ are assumed to be random vectors with Laplacian distribution and diagonal covariance matrices. Let $\sigma_{n_t}^d$ and $\sigma_{y_t}^d$ denote the standard deviations of n_t^d and y_t^d, respectively. The distributions of N_t and Y_t are modeled as

$$\begin{aligned} P(Y_t) &= \prod_{d=1}^{D} P(y_t^d) = \prod_{d=1}^{D} \frac{\alpha_{y_t}^d}{2} e^{-\alpha_{y_t}^d |y_t^d|} \\ P(N_t) &= \prod_{d=1}^{D} P(n_t^d) = \prod_{d=1}^{D} \frac{\alpha_{n_t}^d}{2} e^{-\alpha_{n_t}^d |n_t^d|} \\ \alpha_{y_t}^d &= \frac{\sqrt{2}}{\sigma_{y_t}^d}, \text{ and } \alpha_{n_t}^d = \frac{\sqrt{2}}{\sigma_{n_t}^d} \end{aligned} \tag{4}$$

Since Y_t and N_t are independent,

$$P(X|Y_t) = P(N_t) \tag{5}$$

Consider M reference blocks $Y_m,\ m = 0, \cdots, M-1$. Let m_t be the index of the true predictor. We assume that the true predictor is always one of Y_m, thus $0 \le m_t < M$. $Y_m = \{y_m^d\}_{d=1,\cdots,D}, m \neq m_t$ are identically distributed as Y_t. Let

$$N_m = X - Y_m = Y_{m_t} + N_{m_t} - Y_m,$$

then the standard deviation of N_m is

$$\left(\sigma_n^d\right)^2 = 2\left(\sigma_{y_t}^d\right)^2 + \left(\sigma_{n_t}^d\right)^2 - 2\sigma_{y_t y_m}^d$$

Although in general N_m is not Laplacian distributed, in video coding problems of interest we usually have that $\sigma_{y_t}^d \gg \sigma_{n_t}^d$ and thus N_m can still effectively being modeled as a Laplacian random vector with distribution given by

$$P(N_m) = \prod_{d=1}^{D} P(n_m^d) = \prod_{d=1}^{D} \frac{\alpha_n^d}{2} e^{-\alpha_n^d |n|},\ \alpha_n = \frac{\sqrt{2}}{\sigma_n^d} \tag{6}$$

In our system, we use coset coding and thus the encoder has to select the number of cosets (and thus number of bits) to encode each DCT coefficient. Let $B = \{b^d\}_{d=1,\cdots,D}$ be the bit allocation vector where b^d denotes the number of bits we assign to the d-th DCT coefficient. Then $C = \{c^d\} = \left\{2^{(b^d)}\right\}$, $d = 1, \cdots, D$ denotes the number of cosets to encode the d-th DCT coefficient.

The distortion of the reconstructed frame is defined as

$$\mathcal{D} = E\left[\left(\hat{X}_{m^*} - X\right)^2\right], \tag{7}$$

where m^* is the index selected by the ML algorithm (which need not be the true side information.)

Note that we do not consider the impact of quantization on $\mathcal{D}$ in the analysis. X and Y_m are modeled as continuous random vectors and the coset encoding operation is modeled as a modulo operation, i.e., encoding X using C leads to $\tilde{X} = \mod(X, C)$. This approximation is only used for our analysis; in our experiments, the DCT coefficients are quantized prior to coset coding.

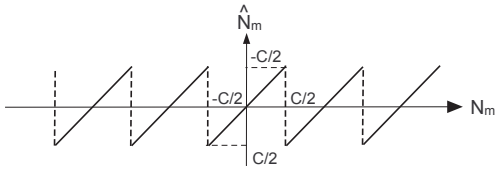

Fig. 2. $\hat{N}_m$ is a periodic function of N_m

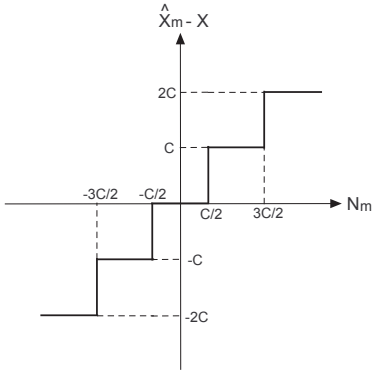

Fig. 3. $\hat{X}_m - X$ as a function of N_m

4. ML DECODING RATE-DISTORTION ANALYSIS

We start from the simplest case, where $D = 1$ and $M = 1$, and then extend the analysis to a more general case where $D > 1$ and $M \to \infty$.

4.1. Coset Coding and ML Decoding

Define $\hat{N}_m = \hat{X}_m - Y_m$. Given (4) and (5),

$$L(\hat{X}_m, Y_m) = \frac{\alpha_n}{2} e^{-\alpha_n |\hat{N}_m|} \tag{8}$$

Since $L(\hat{X}_m, Y_m)$ is a monotonically decreasing function of $|\hat{N}_m|$, ML decoding is equivalent to finding the minimum $|\hat{N}_m|$[1].

Coset coding introduces a "modulo effect" in the likelihood computation since the encoding is ambiguous and we have that $|\hat{N}_m| \leq \frac{C}{2}$ at the decoder. $\hat{N}_m$ is a periodic function of N_m.

$$\hat{N}_m = g(N_m) = \mod (N_m + \frac{C}{2}, C) - \frac{C}{2} \tag{9}$$

as depicted in Figure 2. Since $\hat{X}_m - X = \hat{N}_m - N_m$, a decoding error occurs when $|N_m| > \frac{C}{2}$. We can represent the decoding error as a function of N_m, as depicted in Figure 3.

When $M = 1$ (i.e., the true predictor is given):

$$\mathcal{D} = 2 \sum_{i=1}^{\infty} (iC)^2 \int_{\frac{2i-1}{2}C}^{\frac{2i+1}{2}C} \frac{\alpha_{n_t}}{2} e^{-\alpha_{n_t} |n_t|} dn_t. \tag{10}$$

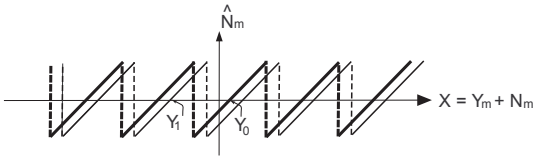

Fig. 4. Illustration of the Distortion Analysis when $D = 1$ and $M = 2$

4.2. Two References

Now consider ML decoding when $M = 2$ references Y_0 and Y_1 are available. Without loss of generality assume that Y_0 is the true predictor. ML decoding can then been seen to be comparing $g(X - Y_0)$ and $g(X - Y_1)$, as illustrated in Figure 4. Since $\sigma_n \gg \sigma_{n_t}$, on average the MSE introduced when a false predictor is picked is much higher than when a true predictor is picked. Hence the overall distortion increases when we have multiple references.

4.3. Asymptotic Behavior ($M \to \infty$)

In typical video coding scenarios, M can be a very large number, e.g., a full search motion estimation algorithm usually would choose the best predictor from more than 1000 candidates ($M > 1000$). This is the motivation to study the asymptotic behavior of $\mathcal{D}$ as M increases.

The probability of the true predictor Y_{m_t} being picked at the decoder can be expressed as

$$P(m^* = m_t) = P(|\hat{N}_{m_t}| < \min_{m \neq m_t} (|\hat{N}_m|)) \tag{11}$$

As $M \to \infty$, $\min_{m \neq m_t}(|\hat{N}_m|) \to 0$, hence $P(m^* = m_t)$ tends to 0 as M increases. Thus we can assume that a false predictor is likely to be picked at the decoder [2]. We further simplify the problem by assuming

$$\hat{N}_{m^*} = \min(\hat{N}_m) = 0$$

Now $\mathcal{D}$ can be computed as

$$\begin{aligned} \mathcal{D} &= \sum_{i=1}^{\infty} (iC)^2 \, P\left(N_{m^*} = iC | \hat{N}_{m*} = 0\right) \\ &= 2 \sum_{i=1}^{\infty} (iC)^2 \frac{\frac{\alpha_n}{2} e^{-\alpha_n |iC|}}{\sum_{j=-\infty}^{\infty} \frac{\alpha_n}{2} e^{-\alpha_n |jC|}} \end{aligned} \tag{12}$$

4.4. Multidimensional DCT Vector ($D > 1$)

When $D > 1$, we can follow the same strategy as in Section 4.2 to compute the expected distortion. However, as D increases, the computation can become intractable due to the increasing difficulty of representing regions corresponding to different distortions. We need to represent different regions corresponding to

$$\left\{ \hat{x}_m^d - x^d \right\}_{d=1,\cdots,D} = \left\{ k^d c^d \right\}_{d=1,\cdots,D},$$

[1] This is not true when $D > 1$, since usually $\alpha_n^d \neq \alpha_n^k$, when $d \neq k$

[2] $\hat{X}$ can be correctly decoded even if a false predictor Y_m is picked. The criterion for correct decoding is $|X_{m^*} - Y_{m^*}| \leq \frac{C}{2}$, not $m^* = m_t$

where $k^d \in \mathbb{N}$. As D increases, the number of combinations of $\{k^d\}_{d=1,\cdots,D}$ increases beyond what would be computationally manageable. Thus, instead of considering all dimensions jointly for the likelihood test, we simplify the computation by considering each dimension individually for the likelihood test and in the distortion computation. The total expected distortion is approximated as the sum of all expected distortions of each dimension computed independently.

Thus, in the general case where $D > 1$ and $M \to \infty$, the total MSE can be expressed as

$$\mathcal{D} \approx 2 \sum_{d=1}^{D} \sum_{i=1}^{\infty} \left(ic^d\right)^2 \frac{\frac{\alpha_n^d}{2} e^{-\alpha_n^d |ic^d|}}{\sum_{j=-\infty}^{\infty} \frac{\alpha_n^d}{2} e^{-\alpha_n^d |jc^d|}} \qquad (13)$$

4.5. Bit Allocation

Since here only integer bits are assigned to each DCT coefficient, we can easily use the Viterbi algorithm to find the bit allocation vector B to minimize (13) under the rate constraint $\sum_{d=1}^{D} b^d \leq B_T$, where B_T is the total number of bits per block.

5. EXPERIMENTAL RESULTS

In our experiments the size of the block is 8×8 (i.e. $D = 64$) and 8×8 DCT is used. All DCT coefficients are quantized with the same uniform scalar quantizer. The test video sequence is "Foreman" in CIF format. We encode the first 20 frames in "I-P" mode to gather information to initialize the correlation model. For each block in P-frame, we consider the blocks from the previous I-frame with minimum SAD as the true reference block. $\sigma_{n_t}^d$, $\sigma_{y_t}^d$ and σ_n^d are estimated based on the collected data. The remaining frames are coded as "I-P-I-S-I-S-I-S-I-S", where "S" denotes the Wyner-Ziv coded frame. The correlation model is updated every 10 frame using the new collected data from the P-frame. In S-frames, the difference of the DC coefficients are entropy coded, while the remaining AC coefficients are coset coded. EOB information is sent before the AC coefficients. We allow blocks in S-frames to be intra-coded if the intra-coded rate is lower than coset-coded rate. In total 100 frames are coded. Currently the bit allocation is decided based on the statistic model at the 20th frame (i.e. the initial model), adaptive bit allocation is left for future work. R-D data shown is for the S-frames only.

To demonstrate the efficacy of the proposed methods, the ML motion estimation method is compared against the CRC based motion estimation used in PRISM [3], and the proposed RD optimized bit allocation is compared against a bit allocation scheme designed based on DISCUS [5]. The DISCUS based bit allocation scheme is designed as follows: First the minimum distance Q_D between 2 symbols in a coset so to achieve a target decoding error probability is computed [5]. Assume the quantization step size used in our system is Q, then

$$B = \left\lfloor \log_2 \left(\frac{Q_D}{Q} \right) \right\rfloor$$

We show the experimental results of 4 different systems: (i) CRC-16: 16-bit CRC motion estimation and DISCUS based bit allocation[3]. (ii) CRC-16-EOB: 16-bit CRC motion estimation, DISCUS based bit allocation and EOB. (iii) ML-EOB: ML motion estimation, DISCUS based bit allocation and EOB. And (iv) ML-EOB-RD: ML motion estimation, proposed RD optimized bit allocation scheme.

[3]Note that here only the lower frequency coefficients are coset coded while the higher frequency coefficients are entropy coded, the same as the original PRISM system.

Figure 5 shows the experimental results. All the systems coded with the EOB approach outperform "CRC-16" by large margin. On average "ML-EOB-RD" shows around 1dB improvement against "ML-EOB", which validates the rate-distortion analysis and the bit allocation strategy.

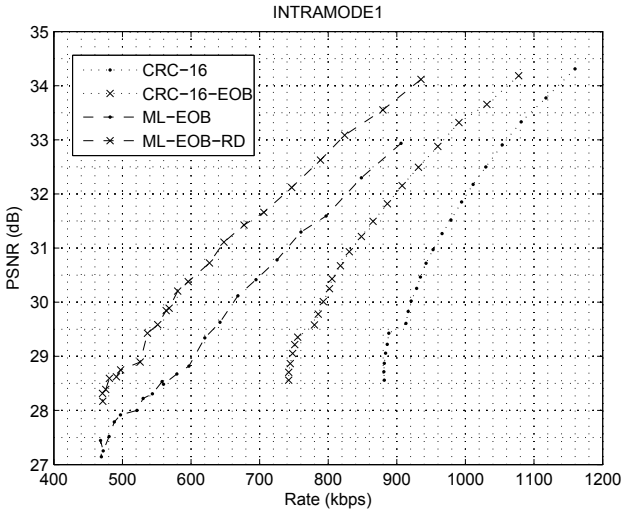

Fig. 5. PSNR curve of various systems described in Section 5.

6. REFERENCES

[1] N.-M. Cheung, C. Tang, A. Ortega, and C. S. Raghavendra, "Efficient Wavelet-based Predictive Slepian-Wolf Coding for Hyperspectral Imagery," *EURASIP Journal on Signal Processing -Special Issue on Distributed Source Coding*, vol. 86, no. 11, pp. 3180–3195, 2006.

[2] A. Aaron, D. Varodayan, and B. Girod, "Wyner-Ziv Residual Coding of Video," in *Proc. Picture Coding Symposium*, Beijing, China, April 2006.

[3] R. Puri and K. Ramchandran, "PRISM: A video coding architecture based on distributed compression principles," Tech. Rep., University of California, Berkeley, 2002.

[4] I. H. Tseng and A. Ortega, "Motion estimation at the decoder using maximum likelihood techniques for distributed video coding," in *Proc. Asilomar Conference on Signals and Systems*, Pacific Grove, CA, November 2005.

[5] S. S. Pradhan and K. Ramchandran, "Distributed Source Coding Using Syndromes (DISCUS): Design and Construction," *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 626–643, March 2003.