Title: Correlation modeling with decoder-side quantization distortion estimation for distributed video coding

Authors: Jozef Škorupa, Jan De Cock, Jürgen Slowack, Stefaan Mys, Nikos Deligiannis, Peter Lambert, Adrian Munteanu and Rik Van de Walle

In: Proceedings of Picture Coding Symposium, December 2010.

CORRELATION MODELING WITH DECODER-SIDE QUANTIZATION DISTORTION ESTIMATION FOR DISTRIBUTED VIDEO CODING

Jozef Škorupa, Jan De Cock, Jürgen Slowack, Stefaan Mys, Peter Lambert, Rik Van de Walle

Ghent University – IBBT,
ELIS – Multimedia Lab,
Gaston Crommenlaan 8 bus 201,
B-9050 Ledeberg-Ghent, Belgium

Nikos Deligiannis and Adrian Munteanu

Vrije Universiteit Brussel – IBBT
Electronics and Informatics Department
Pleinlaan 2, B-1050 Brussels, Belgium

ABSTRACT

Aiming for low-complexity encoding, distributed video coders still fail to achieve the performance of current industrial standards for video coding. One of most important problems in this area is the accurate modeling of the correlation between the predicted signal and the original video. In our previous work we showed that exploiting the quantization distortion can significantly improve the accuracy of a correlation estimator. In this paper we describe how the quantization distortion can be exploited purely at the decoder side without any performance penalty when compared to an encoder-aided system. As a result, the proposed correlation estimator delivers state-of-the-art modeling accuracy while neatly fitting the low-encoder-complexity characteristic of distributed video coding.

Index Terms— distributed video coding, correlation modeling, quantization distortion

1. INTRODUCTION

Based on Wyner-Ziv theory, distributed video coders represent an alternative to the fixed complexity distribution in predictive coders, specified by the current industrial standards. Offering low-complexity encoding, distributed coders are ideally suited for application scenarios involving deployment of a large number of power- or hardware-constrained devices, e.g., (wireless) video surveillance.

Unfortunately, Distributed Video Coding (DVC) still fails to match the coding performance of predictive coding schemes. One of the most important factors [1] influencing the coding performance is the accurate modeling of the correlation between the original video signal available at the encoder and the predicted signal (or so-called side information), which is generated at the decoder in DVC systems.

Typical improvements in correlation modeling result in gains which are limited to high rates [2, 3]. Recently, we have proposed a correlation estimator which exploits the distortion induced by quantization in order to achieve systematic coding gains throughout the entire rate region [4, 5]. The proposed method employed the encoder to estimate the distortion and send necessary information through the video stream. This introduces an unnecessary, albeit small, overhead both in terms of bit-rate and encoder complexity. To alleviate this problem, we have designed a pure decoder-based method to estimate the quantization distortion which is employed in our correlation estimator.

The main contribution of this paper is the description of a decoder-based correlation estimator which exploits the quantization distortion in order to achieve an accurate estimation of the correlation throughout the entire rate region. We describe our correlation estimator itself in Sect. 3.1, while in Sect. 3.2 we focus on the decoder-based estimation of quantization distortion.

In order to assess the impact of pure decoder-based estimation as opposed to encoder-aided estimation, and to position our estimator relative to state-of-the-art correlation estimators, we have conducted experiments using the DVC architecture described in Sect. 2. The results reported in Sect. 4 indicate that exploiting the quantization distortion in correlation estimation yields significant coding gains, while doing that purely at the decoder brings no performance penalty compared to encoder-aided estimation.

2. CODEC ARCHITECTURE

The general architecture of our codec (Fig. 1) is based on the common DVC schema well described in the literature, e.g. [6]. The frames are divided into key frames I_n and Wyner-Ziv frames X_n based on a fixed GOP structure. Key frames are intra coded using an H.264/AVC intra coder. For each Wyner-Ziv frame X_n, the decoder generates a prediction Y_n as a motion-compensated interpolation of previously decoded frames [6]. The prediction can be seen as a noisy version of the original frame corrupted by virtual noise[1] $N_n = X_n - Y_n$. By applying an error-correcting code at the encoder and sending parity bits to the decoder, the errors in the prediction are

[1] Due to the description of the problem using the virtual noise abstraction, correlation estimation is often referred to as virtual-noise estimation.

corrected and the decoded Wyner-Ziv frame $\hat{X}_n$ is obtained[2].

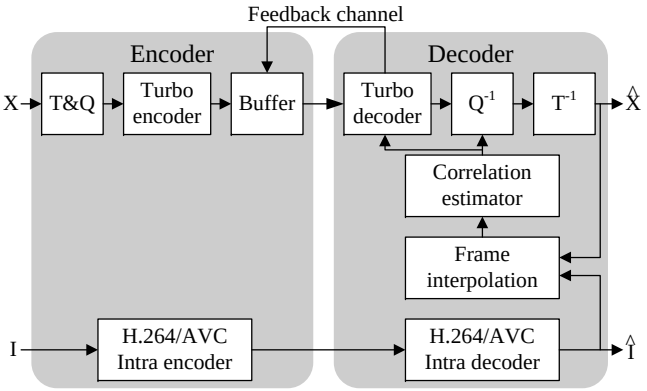

Fig. 1. The codec extracts bitplanes from transformed (T) and quantized (Q) frames, and the parity bits calculated for each bitplane by a turbo coder are used to correct the prediction at the decoder.

3. CORRELATION ESTIMATOR

In this section we describe our correlation estimator in two steps. In Sect. 3.1, we describe correlation estimation as proposed in our previous work [5]. Then we show how this estimation can be improved by a no-reference quality assessment at the decoder (Sect. 3.2).

3.1. Correlation estimation

As common in the DVC literature, we will assume that the virtual noise follows a Laplacian distribution, given by

$$f_{X|Y}(x|y) = \frac{\lambda_k}{2}\, e^{-\lambda_k|x-y|} \;, \tag{1}$$

where $\lambda_k > 0$ is the scale parameter, which is to be estimated by the correlation estimator on a block basis.

Most of the state-of-the-art estimators rely on the motion-compensated residual frame R to estimate the correlation. The residual frame is defined as

$$R_n(\boldsymbol{p}) = \hat{X}_{n-i}(\boldsymbol{p} + \boldsymbol{m}^-(\boldsymbol{p})) - \hat{X}_{n+i}(\boldsymbol{p} + \boldsymbol{m}^+(\boldsymbol{p})) \;, \tag{2}$$

where $\boldsymbol{p}$ denotes the position in the frame and $\boldsymbol{m}^-(\boldsymbol{p}))$, $\boldsymbol{m}^+(\boldsymbol{p}))$ denote the motion vectors pointing to the previous reference frame $\hat{X}_{n-i}$ and the future reference frame $\hat{X}_{n+i}$, respectively.

As we have pointed out in our previous work [4, 5], the residual frame alone is not enough to obtain an accurate correlation estimation for the following reasons. Firstly, the residual frame does not properly reflect changes in quantization settings. Secondly, the spatial properties of the residual frame usually differ from those of the virtual-noise frame, which impairs the correlation estimation in the frequency domain when the transformed residual frame is used. To solve these issues, we have proposed a two-tier estimator [5], as described in the following.

In the first step, the virtual noise in the pixel domain is estimated using maximum-likelihood estimation

$$\lambda_k = \frac{1}{n_k} \;, \tag{3}$$

where n_k is the average of the absolute value of the noise in a block B_k. Because the value of n_k is not known at the decoder, it is estimated using

$$n_k = \frac{r_k}{2} + q \;, \tag{4}$$

where r_k is the average of the absolute value of the residual frame in the block B_k and q is the average (absolute) distortion of the reference frames. Eq. (4) is based on an analytical model for motion-compensated interpolation we described in [5]. The rationale behind the model is that the virtual noise is a compound distortion consisting of a mismatch caused by block-based motion estimation ($r_k/2$) and the distortion already present in the reference frames (q).

In our previous work, we approximated the second term q with the quantization distortion in the intra coded frames, which was calculated at the encoder side and sent along the stream. The main contribution of this paper is a method for estimation of q at the decoder side, which we describe in Sect. 3.2.

The second tier of our correlation estimator employs the virtual noise estimated in the pixel domain to obtain estimation of the correlation in the frequency domain. As we already pointed out in [5], this approach has the advantage of increased statistical support for the estimation and alleviates the problem of improper spatial properties of the residual frame.

As pointed out by several authors [3, 5], the per-band variance of the noise signal in the frequency domain can be expressed using the variance of the noise in the pixel domain as

$$\sigma^2_{N_k^{(b)}} = \sigma^2_{N_k}\, V^{(b)} \;, \tag{5}$$

where $\sigma^2_{N_k}$ is variance of the pixel domain noise which is calculated as $\sigma^2_{N_k} = 2/\lambda_k^2$ using parameter λ_k obtained in the pixel domain, and $V^{(b)}$ is a scale factor for band b, which depends on the autocorrelation of the noise signal. In our previous work, we modeled the autocorrelation of the noise as a mixture of autocorrelation of the previously decoded Wyner-Ziv frame, estimated at the decoder, and the autocorrelation of the previously decoded key frame. The latter was estimated at the encoder and communicated through the bitstream. In this paper we show how $V^{(b)}$ can be accurately estimated purely at the decoder, alleviating the need for encoder-decoder communication (Sect. 3.2).

[2] In the following, we drop the temporal index of the frame when it is clear from the context.

3.2. Decoder-side distortion estimation

The distortion of the key frames and their spatial properties (autocorrelation) exploited in our estimator (Sect. 3.1) are very important for accurate correlation estimation, especially at low bit rates. Albeit very small, the approach from [5] still imposes overhead on the bit rate and encoder complexity by measuring the required values at the encoder and sending them along the stream. We now propose a method to estimate q and $V^{(b)}$ required in (4) and (5), respectively, purely at the decoder side.

The method is based on no-reference image quality assessment proposed by Brandão and Queluz [7]. The distribution in frequency band b of the signal to be coded—in the case of the H.264/AVC intra coder, the residual after subtracting the intra prediction—is assumed to be Laplacian with scale parameter $\alpha^{(b)}$. For each band the scale parameter is obtained as a weighted average of the maximum-likelihood estimation $\alpha_{ML}^{(b)}$ and prediction from the previously estimated bands $\alpha_p^{(b)}$ as follows

$$\alpha^{(b)} = (1 - r_0^{(b)})\alpha_{ML}^{(b)} + r_0^{(b)}\alpha_p^{(b)} \ . \tag{6}$$

In (6), the weighting coefficient $r_0^{(b)} = M_0^{(b)}/M$ is taken as the ratio between the number of coefficients in band b quantized to zero $M_0^{(b)}$ and the total number of coefficients in the band M. The prediction term $\alpha_p^{(b)}$ in (6) is used to compensate for overflow in maximum-likelihood estimation when almost all coefficients are quantized to zero $M_0^{(b)} \to M$, i.e. at low bit rates and/or in high frequency bands.

The maximum-likelihood estimation operates with the number of zero coefficients $M_0^{(b)}$ and the sum of absolute values of the inverse quantized coefficients $S^{(b)} = \sum_{l=0}^{M-1} |X_l^{(b)}|$, and the estimated parameter reads as

$$\alpha_{ML}^{(b)} = -\frac{2}{Q} \log \frac{-M_0^{(b)}Q + \sqrt{D}}{2MQ + 4S^{(b)}} \tag{7}$$

where Q denotes the quantizer step size and D reads as $D = (M_0^{(b)}Q)^2 - 4(MQ + 2S^{(b)})((M - M_0^{(b)})Q - 2S^{(b)})$.

For all AC bands, $\alpha_p^{(b)}$ is predicted from already estimated parameters (estimation is performed in zig-zag scan order) as

$$\alpha_p^{(b)} = \boldsymbol{\alpha^{(b)}}\boldsymbol{\beta^{(b)}}^T \ , \tag{8}$$

where $\boldsymbol{\alpha^{(b)}} = (1, \alpha^{(0)}, \dots \alpha^{(b-1)})$ and $\boldsymbol{\beta^{(b)}}$ is the predictor vector. $\boldsymbol{\beta^{(b)}}$ is trained offline by minimizing the prediction error on a given training set. The training procedure is described in great detail in [7].

Knowing the parameter of the distribution, one can calculate the squared distortion that a scalar quantizer with known quantization step and deadzone parameter will induce on such signal. Obtaining the distortion includes some heavy, albeit straightforward, algebraic manipulations [8]. We will denote with $\sigma_{(b)}^2$ the squared distortion induced by quantization of the band b. Recognizing that the DCT is an orthogonal transform[3], we can clearly see that the squared distortion in the key frame, i.e. variance of the frame difference $I - \hat{I}$, could be expressed (for a 4×4 transform) as

$$\sigma_{I-\hat{I}}^2 = \frac{1}{16} \sum_{b=0}^{15} \sigma_{(b)}^2 \ . \tag{9}$$

Assuming that $I - \hat{I}$ follows a Laplace distribution, we can express the required average absolute value of the distortion in the key frames as

$$q = \sqrt{\frac{\sigma_{I-\hat{I}}^2}{2}} \ , \tag{10}$$

which can be directly used in (4).

Having estimated the quantization distortion q, we still need to obtain the scaling factor $V^{(b)}$ from (5). $V^{(b)}$ can be expressed as

$$V^{(b)} = \sum_{\boldsymbol{p_1}} \sum_{\boldsymbol{p_2}} c_{\boldsymbol{p_1}}^{(b)} c_{\boldsymbol{p_2}}^{(b)} R_N(\boldsymbol{p_1}, \boldsymbol{p_2}) \tag{11}$$

where $c_{\boldsymbol{p_1}}^{(b)}$ and $c_{\boldsymbol{p_2}}^{(b)}$ are the corresponding coefficients of the transformation [3, 5], $R_N(\boldsymbol{p_1}, \boldsymbol{p_2})$ is the autocorrelation of $N = X - Y$, and the pixel positions $\boldsymbol{p_1}, \boldsymbol{p_2}$ loop through the whole (4×4) transform block. We proposed to approximate the autocorrelation function as a weighted average of two autocorrelations—autocorrelation of frame difference between the previously decoded Wyner-Ziv frame and its prediction $\hat{X}_p - Y_p$, and autocorrelation of frame difference between the last decoded key frame and its original $I - \hat{I}$. Together with the weighting coefficient, the autocorrelation reads as

$$R_N(.) = wR_{\hat{X}_p - Y_p}(.) + (1 - w)R_{I-\hat{I}}(.) \tag{12}$$

$$w = \sigma_{\hat{X}_p - Y_p}^2 / (\sigma_{\hat{X}_p - Y_p}^2 + \sigma_{I-\hat{I}}^2) \ . \tag{13}$$

Analyzing (11) and (12), one can clearly see that the scaling factor $V^{(b)}$ can be expressed as the following average: $V^{(b)} = wV_{\hat{X}_p - Y_p}^{(b)} + (1 - w)V_{I-\hat{I}}^{(b)}$, where both terms $V_{\hat{X}_p - Y_p}^{(b)}$ and $V_{I-\hat{I}}^{(b)}$ are given as a sum similar to (11) with respective autocorrelation functions $R_{\hat{X}_p - Y_p}$ and $R_{I-\hat{I}}$. Because $\hat{X}_p - Y_p$ can be constructed at the decoder, $\sigma_{\hat{X}_p - Y_p}^2$ and $R_{\hat{X}_p - Y_p}$ can be calculated directly. $\sigma_{I-\hat{I}}^2$ is obtained by no-reference distortion estimation (9) and we propose to estimate the scaling factor $V_{I-\hat{I}}^{(b)}$ with the already obtained information as

$$V_{I-\hat{I}}^{(b)} = \sigma_{(b)}^2 / \sigma_{I-\hat{I}}^2 \ . \tag{14}$$

Using the method described in the previous paragraphs, the proposed estimator alleviates the need for encoder-decoder communication, saving the bit rate and encoder complexity required for such communication.

[3] In case of the H.264/AVC integer transform, one needs to account for the scaling first.

Table 1. The relative error $|q - q_e|/q_e$ of no-reference distortion assessment according to Eq. (10).

QP	M&D	Foreman	Coastguard	Mobile
20	0.04	0.03	0.12	0.04
25	0.09	0.01	0.15	0.02
30	0.07	0.04	0.17	0.01
35	0.03	0.03	0.23	0.03

4. EXPERIMENTAL RESULTS

We trained the predictor vector (8) on 12 sequences, different than the test sequences. Tests were conducted on four CIF sequences (30 Hz): Mother and daughter (M&D), Foreman, Coastguard and Mobile. GOP sizes of 4 and 8 were used.

First, we explored the accuracy of the distortion q estimated at the decoder (10) by comparing it to the distortion q_e as measured at the encoder. The results (Tab. 1) show that the proposed no-reference quality assessment provides a highly accurate estimation for most sequences.

To assess the impact of the correlation estimator on the RD performance, we have compared our DVC system employing the proposed estimator, the estimator proposed in our previous work [5] and the state-of-the-art correlation estimator as described by Brites and Pereira in [1]. Our estimator outperforms the state-of-the-art with average bit rate gains of 2.8% (Mobile), 10.8% (Foreman), 13.2% (Coastguard) and 13.9% (Mother and daughter), as measured by the Bjøntegaard-Delta metric. These gains are systematic over the entire rate region and are brought by the use of the quantization term q in (4); we notice also that having gains over the entire rate range is in steep contrast to typical gains reported in literature, which tend to diminish towards low rates.

Comparing to our previous work [5], where the quantization distortion is calculated at the encoder, we can see that (i) no-reference assessment at the decoder provides an accurate estimation of the distortion and (ii) the proposed estimator is not hindered by the lack of exact data from the encoder. On the contrary, by eliminating the need for sending the distortion values in the stream, small gains—between 0.2% and 0.5% of Bjøntegaard-Delta rate—are achieved. Moreover, the estimator seems to be robust against occasional inaccuracies in the distortion assessment as shown in Fig. 2, where the sequence with the largest relative error is depicted.

5. CONCLUSIONS

It is clear (from results presented here or in our previous work [4, 5]) that exploiting the distortion in the reference frames significantly improves the accuracy of correlation estimation in DVC. The method proposed in this paper estimates this distortion purely at the decoder side. We have shown that, when compared to encoder-aided estimation, using no-reference quality assessment does not decrease the accuracy of the correlation estimator while it reduces the bit rate and encoder complexity.

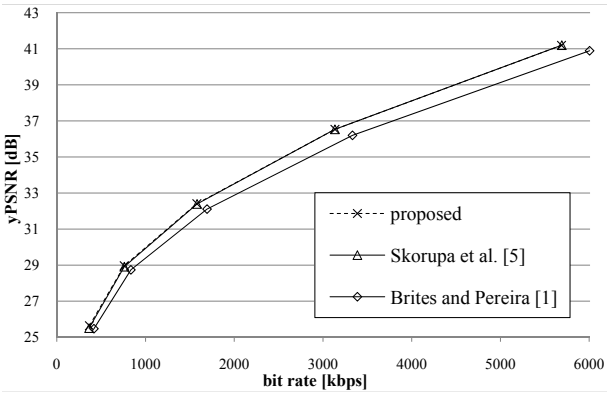

Fig. 2. Exploiting the quantization distortion leads to significant gains over the state-of-the-art estimators, whether the no-reference decoder-side assessment (proposed) or encoder-aided exact measurement ([5]) is employed. Coastguard, CIF@30Hz, GOP 8.

6. REFERENCES

[1] C. Brites and F. Pereira, "Correlation noise modeling for efficient pixel and transform domain Wyner-Ziv video coding," *IEEE Trans. Circuits Syst. for Video Technol.*, vol. 18, no. 9, pp. 1177–1190, Sep 2008.

[2] X. Huang and S. Forchhammer, "Improved virtual channel noise model for transform domain Wyner-Ziv video coding," in *Proc. IEEE Inter. Conf. on Acoust., Speech, Signal Processing*, Apr 2009.

[3] X. Fan, O. C. Au, and N. M. Cheung, "Adaptive correlation estimation for general Wyner-Ziv video coding," in *Proc. IEEE Inter. Conf. Image Processing*, Nov 2009.

[4] J. Slowack, S. Mys, J. Škorupa, P. Lambert, and R. Van de Walle, "Accounting for quantization noise in online correlation noise estimation for distributed video coding," in *Proc. Picture Coding Symposium*, May 2009.

[5] J. Škorupa, J. Slowack, S. Mys, N. Deligiannis, J. De Cock, P. Lambert, A. Munteanu, and R. Van de Walle, "Exploiting quantization and spatial correlation in virtual-noise modeling for distributed video coding," *Signal Processing: Image Commun.*, vol. 25, no. 9, pp. 674–686, Oct 2010.

[6] C. Brites, J. Ascenso, J. Quintas Pedro, and F. Pereira, "Evaluating a feedback channel based transform domain Wyner-Ziv video codec," *Signal Processing: Image Commun.*, vol. 23, no. 4, pp. 269–297, Apr 2008.

[7] T. Brandão and M. P. Queluz, "No-reference image quality assessment based on dct domain statistics," *Signal Process.*, vol. 88, no. 4, pp. 822–833, Apr 2008.

[8] G. J. Sullivan, "Efficient scalar quantization of exponential and Laplacian random variables," *IEEE Trans. Inform. Theory*, vol. 42, no. 5, pp. 1365–1374, Sep 1996.