STUDYING TEMPORAL CORRELATION NOISE MODELING FOR PIXEL BASED WYNER-ZIV VIDEO CODING

Catarina Brites[1], João Ascenso[2], Fernando Pereira[3]

[1]catarina.brites@lx.it.pt, [2]joao.ascenso@lx.it.pt, [3]fp@lx.it.pt
[1,3]Instituto Superior Técnico – Instituto de Telecomunicações
[2]Instituto Superior de Engenharia de Lisboa – Instituto de Telecomunicações

ABSTRACT

Wyner-Ziv (WZ) video coding – a particular case of distributed video coding (DVC) – is a new video coding paradigm based on two major Information Theory results: the Slepian-Wolf and Wyner-Ziv theorems. Recently, practical WZ video coding solutions were proposed with promising results. Most of the solutions available in the literature, model the correlation noise between the original frame and the so-called side information by a given distribution whose relevant parameters are estimated in an offline process, at the encoder. In this paper, three algorithms are proposed towards a more realistic WZ coding approach by performing online estimation of the error distribution at the decoder. Both algorithms explore temporal correlation between frames however with different levels of granularity: frame, block and pixel levels; better rate-distortion (RD) performance is achieved for lower granularity (pixel) level.

1. INTRODUCTION

Nowadays, the digital video coding solutions available rely on the powerful hybrid block-based motion compensation and DCT transform (MC/DCT) architecture. All the ITU-T VCEG and ISO/IEC MPEG standards follow this approach, mostly targeting applications where the video content is encoded once and decoded multiple times, e.g. broadcasting or video streaming. In such applications, the video codec architecture is primarily driven by the one-to-many model of a single complex encoder and multiple light (cheap) decoders; typically the encoder is 5 to 10 times more complex than the decoder [1]. The complexity burden of the encoder is mainly associated with the motion estimation/ compensation task, which is the major responsible for the high rate-distortion performance achieved.

However, this architecture is being challenged by several emerging applications such as wireless video surveillance, multimedia sensor networks, wireless PC cameras and mobile camera phones. These applications have different requirements from those targeted by traditional video delivery systems, e.g. in wireless video surveillance systems, low cost encoders are important since there is a high number of encoders and only one or few decoders.

Distributed video coding fits well in these scenarios, since it enables to explore the video statistics, partially or totally, at the decoder only, relying on a low encoding complexity. From the Information Theory, the Slepian-Wolf theorem [2] states that it is possible to compress two statistically dependent signals, X and Y, in a distributed way (separate encoding, jointly decoding) using a rate similar to that used in a system where the signals are encoded and decoded together, i.e. like in traditional video coding schemes. The complement of Slepian-Wolf coding for lossy compression is Wyner-Ziv coding [3]. The WZ coding deals with lossy source coding of X with side information Y at the decoder and enables a flexible allocation of complexity between the encoder and the decoder. The side information is usually interpreted as an attempt made by the decoder to obtain an estimate of the original frame. In the WZ coding scenario, error correcting codes are used to improve the quality of the side information until a target quality for the final decoded frame is achieved.

One of the most interesting DVC approaches is the turbo-based pixel domain Wyner-Ziv coding scheme presented in [4], where the decoder is responsible to explore all the source statistics, and therefore to achieve compression following the Wyner-Ziv paradigm. Since DVC implies a statistical mind set, the coding efficiency of Wyner-Ziv coding solutions depends critically on the capability to model the statistical dependency between the original information at the encoder and the side information computed at the decoder. This is a complex task since the original information is not available at the decoder and the side information quality varies along the sequence, i.e. the error distribution is not temporally constant. For example, when high motion occurs in a sequence is more difficult to predict the Wyner-Ziv frame and the errors in the side information increase significantly. In this paper, new methods are proposed to estimate the correlation noise model based on temporal correlation information, in this case the motion compensated residual obtained at the decoder. The first method proposed models the correlation noise distribution, at the decoder, adaptively at the frame level.

However, the correlation noise statistics between the original and side information frames are not spatially stationary. Usually, the noise or error $N = X - Y$ is estimated without taking into account spatial dependencies. This is an unrealistic assumption because the quality of the side information is not constant across the whole frame due to occlusions or illumination changes. This paper departs from the spatial stationary assumption by proposing a new algorithm performed at the decoder which adapts the correlation noise statistics locally on a block-by-block basis and at the pixel level.

This paper is organized as follows: Section 2 presents a brief summary of the IST-PDWZ codec. In Section 3, offline (encoder-generated) correlation noise statistics models are briefly described to introduce the online (decoder-generated) correlation noise statistics

* The work presented here was developed within DISCOVER, a European Project (http://www.discoverdvc.org), funded under the European Commission IST FP6 programme.

models presented in Section 4, which constitutes the novel element of this paper. Conclusions and some future work topics are presented in Section 5.

2. THE IST-PIXEL DOMAIN WYNER-ZIV (IST-PDWZ) VIDEO CODEC

Figure 1 illustrates the architecture of the IST-PDWZ video codec proposed in [5]; this codec is an improved PDWZ video coding solution which follows the same architecture as the one proposed by Aaron et al. in [4]. There are however some major differences regarding the coding solution proposed in [5], notably a more efficient side information generation scheme at the decoder by using motion compensated frame interpolation with spatial motion smoothing (for more details the reader should consult [6]).

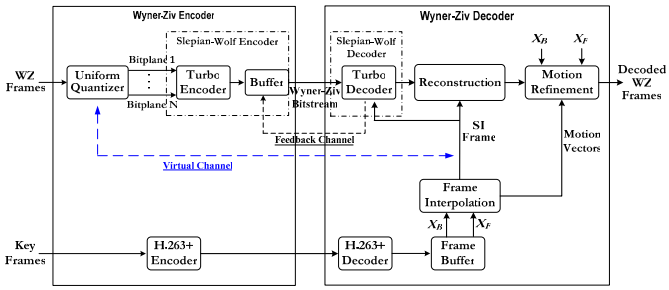

Figure 1 – IST-PDWZ video codec architecture.

In a nutshell, the coding process is as follows: the video frames are organized into key frames and Wyner-Ziv (WZ) frames. The key frames are traditionally intraframe coded. The Wyner-Ziv frame pixel values are quantized using a 2^M-level uniform scalar quantizer. Over the resulting quantized symbol stream, bitplane extraction is performed; each bitplane is then independently turbo encoded and the parity bits are stored in the buffer and transmitted in small amounts upon decoder request via the feedback channel.

At the decoder, the frame interpolation module is used to generate the side information frame, an estimate of the WZ frame, based on previously decoded frames, X_B and X_F. For a Group Of Pictures (GOP) length of 2, X_B and X_F are the previous and the next temporally adjacent key frames.

The side information SI is then used by an iterative turbo decoder to obtain the decoded quantized symbol stream. The turbo decoder is constituted by two soft-input soft-output (SISO) decoders; each SISO decoder is implemented using the Maximum *A Posteriori* (MAP) algorithm. It is assumed the decoder has ideal error detection capabilities, i.e. the turbo decoder is able to measure in a perfect way the current bitplane error probability P_e. For example, if $P_e > 10^{-3}$, the decoder requests for more parity bits from the encoder via feedback channel; otherwise, the bitplane turbo decoding task is considered successful. The side information is also used in the reconstruction module, together with the decoded quantized symbol stream, to help in the WZ frame reconstruction task. The motion refinement module is used to improve the quality of the reconstructed image for a certain bitrate, i.e. after decoding an integer number of bitplanes.

3. OFFLINE (ENCODER) CORRELATION NOISE MODELS

In order to the decoder make use of the side information, obtained at the decoder by frame interpolation, it needs to have reliable knowledge of the model that characterizes the statistical relation between the SI frame and the original frame. The statistical dependency between these two frames corresponds to a virtual channel (see Figure 1) with an error pattern characterized by some statistical distribution (or model) since the side information may be seen as a 'corrupted' version of the original information. If the model accurately describes the WZ − SI relationship, the coding efficiency is high; however, if this model fails or if there is a significant mismatch between the "true" correlation and the estimated one, it will be observed a coding efficiency loss. In the context of Figure 1, this corresponds to less accurate information at the input of the SISO decoders and it will make the turbo decoder spent more bits in order to correct the same amount of errors. The first step towards this direction is to present a study of offline models computed at the encoder using the original information; this will give insight of the maximum or "ideal" performance (and the importance of the correlation noise model in the overall setting) that can be achieved.

3.1. Sequence level correlation noise estimation

In previous works, e.g. [5], [6], the authors used a Laplacian distribution as in (1) to model the statistical correlation between the original frame and the side information.

$$f(WZ - SI) = \frac{\alpha}{2} e^{\alpha |WZ - SI|} \quad (1)$$

The Laplacian distribution is used to convert the side information (pixel values) into soft-input information needed for turbo decoding. In [5], [6] the Laplacian distribution parameter α, given by

$$\alpha^2 = \frac{2}{\sigma^2}, \quad (2)$$

is computed offline, at the encoder, over the whole video sequence and sent to the decoder before the WZ coding procedure starts. It is then kept constant for the decoding of all WZ frames. In (2), σ^2 is the variance of the residual between the WZ and the SI frames. This α calculation process is however not efficient because it does not exploit the variability of the correlation model along time (changes between frames) and space (between regions of a frame).

3.2. Frame level correlation noise estimation

In order to explore the time variability of the correlation noise model, one possible approach is to calculate the variance of the residual between a WZ frame and the corresponding SI frame, and use (2) to compute the Laplacian parameter. This Laplacian distribution parameter value is then used in the decoding process of that WZ frame.

3.3. Block level correlation noise estimation

Recognizing that within the SI frame coexist regions where the frame interpolation was successful (high correlation) and regions where the interpolation has failed (low correlation) will lead to an adaptation of the correlation noise model at the block level (i.e. to a lower granularity level than the frame level). Usually, the frame interpolation algorithm fails often in regions where a high amount of motion occurred or in uncovered regions and it is quite good in regions where the amount of motion is low, e.g. static background. Since different areas of the SI frame have associated different amounts of interpolation errors, it is expected the correlation noise model between the WZ and the SI frames varies within the frame.

In order to exploit the spatial variability of the correlation noise model, the Laplacian distribution parameter is calculated at the $n \times n$ samples block level α_k; the block size considered is equal to the one

used in the frame interpolation stage in order to more easily detect the interpolation errors. The $n \times n$ block variance σ_k^2 is calculated from:

$$\sigma_k^2 = E\left[(WZ_k - SI_k)^2\right] - \left(E\left[(WZ_k - SI_k)\right]\right)^2 \quad (3)$$

where WZ_k and SI_k represent the k^{th} samples block of the WZ and SI frames respectively and $E[.]$ is the expectation operator. Substituting in (2) σ^2 by σ_k^2 given by (3), the α_k value is obtained; all the samples within the k^{th} block are characterized by the same α_k value.

3.4. Experimental results

Table 1 illustrates the rate and the PSNR results obtained for the first 101 frames of the *Foreman* QCIF sequence according to the number of decoded Wyner-Ziv bitplanes (1, 2, 3 or 4). The test conditions for the frame interpolation and motion refinement modules are the same used in [5]. The key frames are encoded with H.263+ intra with a quantization parameter (QP) equal to 13, 10, 8, 5, respectively, depending on the number of decoded Wyner-Ziv bitplanes; using these QP values for the key frames allows to have almost constant decoded video quality for the full set of frames (key frames and WZ). A GOP size of 2 is considered, i.e. one WZ frame in between two Intra coded frames.

Table 1 – RD results for the sequence, frame and block granularity levels for the *Foreman* QCIF sequence.

Nr. Bitplanes / Granularity	1 Rate [kbps]	1 PSNR [dB]	2 Rate [kbps]	2 PSNR [dB]	3 Rate [kbps]	3 PSNR [dB]	4 Rate [kbps]	4 PSNR [dB]
Sequence level (offline)	364.42	31.59	500.77	33.17	667.37	34.57	1045.9	37.6
Frame level (offline)	362.55	31.59	498.19	33.18	664.32	34.57	1042.38	37.6
Block level (offline)	358.55	31.59	486.9	33.18	644.56	34.58	1015.09	37.61

As expected, the better RD results are obtained for the block level adaptation of the correlation noise model. The coarser sequence level adaptation has the worst results and as more fine adaptation is performed (frame and block levels) the RD results improve. The results obtained also show modest gains in RD performance by adapting the correlation model at a finer granularity level, with a maximum of 30.8 kb/s in the 4th bitplane comparing the sequence and block level adaptations. Further gains are expected if the side information exhibits more abrupt quality variations temporally and/or spatially.

4. ONLINE (DECODER) CORRELATION NOISE MODELS

The offline (encoder) correlation noise calculation process is not acceptable and realistic because it requires the encoder to recreate the side information. Since it is used a motion estimation and compensation algorithm to generate the side information, this task cannot be performed in a low complexity encoder (one of the main targets of distributed video coding).

In this context, the realistic approach is to perform dynamic estimation, at the frame, block or even at the pixel level, of the Laplacian distribution parameter in the decoder, where more computational resources are available according to the DVC paradigm. Moreover, the Laplacian distribution parameter does not have to be transmitted from the encoder to the decoder, typically under error prone conditions.

The major novelty of this paper resides on the proposal and study of new α parameter estimation methods that work efficiently at the decoder. This step represents an important departure from previous work in the literature and can lead to a more practical Wyner-Ziv video coding solution since it is no longer necessary to recreate the side information at the encoder side. Three granularity levels are proposed: the frame, the block and the pixel level; both are evaluated in the architecture described in Section 2 (Figure 1). The novel estimation techniques make use of X_B and X_F frames (where X_B and X_F are previously decoded key frames) along with the motion vectors obtained in the side information generation process.

4.1. Frame level correlation noise estimation

In order to estimate the Laplacian distribution parameter at the decoder, it is necessary to define a metric that expresses the variance between the original and the side information, since the original frame is not available at the decoder. The frame level α estimation technique proposed in this paper can be described in the following steps:

i) *Residual frame generation:* It is first computed the residual frame R, between the motion compensated versions of the frames X_B and X_F as follows:

$$R(x,y) = ABS\left(X_B(x + dx_b, y + dy_b) - X_F(x + dx_f, y + dy_f)\right) \quad (4)$$

The X_B $(x + dx_b, y + dy_b)$ and X_F $(x + dx_f, y + dy_f)$ represent the backward and the forward motion compensated frames, respectively and (x, y) corresponds to the pixel location in the R frame. In (4), (dx_b, dy_b) and (dx_f, dy_f) represent the motion vectors for the X_B and X_F frames, respectively.

ii) *Residual frame variance computation:* The variance of the residual frame is then calculated from:

$$\sigma_R^2 = E\left[R(x,y)^2\right] - \left(E\left[R(x,y)\right]\right)^2 \quad (5)$$

iii) *α parameter estimation:* σ_R^2 is a confidence measure of the SI frame creation process which indicates how good the frame interpolation outcome is; ideally σ_R^2 should be close to the variance of the residual between the WZ and the SI frames. So, it is proposed to use the variance metric defined in (5) as a way to represent the variance between the original information and the side information; the α parameter estimate for each WZ frame is then obtained from (2) by substituting σ^2 by the σ_R^2 obtained from (5).

4.2. Block level correlation noise estimation

As previous results have shown, adapting the Laplacian distribution parameter between the WZ and the SI frames at the block level can improve the IST-PDWZ RD performance when compared to a sequence or frame level approach. In this Section, it is proposed a block level α estimation technique performed at the decoder with the aim of improving the RD performance when compared to the frame granularity level. The block level α estimation approach proposed here can be described in the following steps:

i) ***Residual frame generation:*** The residual frame R between the X_B and X_F, both motion compensated, is firstly computed as described in (4).

ii) ***Residual frame k^{th} block variance computation:*** The k^{th} block variance of the residual frame, $\sigma^2_{R_k}$, can be obtained from (5) however considering the expectation operation over the $n \times n$ samples block of frame R instead of the whole frame.

iii) ***α parameter estimation:*** The α parameter estimate for the k^{th} block of the WZ frame is:

$$\hat{\alpha}_k^2 = \frac{2}{\sigma^2_{R_k}} \times \frac{E\left[R_k(x,y)\right]}{E\left[R(x,y)\right]} \quad (6)$$

In (6), $E\left[R_k(x,y)\right]$ is the expectation operation over the $n \times n$ samples block of frame R and $E\left[R(x,y)\right]$ is the expectation operation over the frame R. Experimentally, (6) has shown to fit well the α_k parameter calculated between corresponding k^{th} samples blocks of the WZ and SI frames (Section 3.3).

4.3. Pixel level correlation noise estimation

The coarse to fine strategy can be pursued even further by estimating the correlation noise model at the pixel level. In this Section, it is proposed a technique, performed at the decoder, to estimate the Laplacian distribution parameter at the pixel level; this technique aims to improve the RD performance when compared to coarser granularity levels (frame and block). The pixel level α estimation approach proposed here can be described in the following steps:

i) ***Residual frame generation:*** The residual frame R between the X_B and X_F, both motion compensated, is firstly computed as described in (4).

ii) ***α parameter estimation:*** The α parameter estimate at the (x, y) pixel location is:

$$\hat{\alpha}^2_{(x,y)} = \frac{2}{R(x,y)} \times \frac{1}{E\left[R(x,y)\right]} \quad (7)$$

In (7), $E\left[R(x,y)\right]$ is the expectation operation over the frame R. Experimentally, (7) has shown to fit well the $\alpha_{(x,y)}$ parameter calculated between co-located pixel values of the WZ and SI frames.

4.4. Experimental results

Table 2 shows the IST-PDWZ RD for the same test conditions as defined in Section 3.4, but now using the online decoder correlation noise estimation models proposed in the previous Sections.

Table 2 – RD results for the frame, block and pixel granularity levels for the *Foreman* QCIF sequence.

Nr. Bitplanes / Granularity	1		2		3		4	
	Rate [kbps]	PSNR [dB]	Rate [kbps]	PSNR [dB]	Rate [kbps]	PSNR [dB]	Rate [kbps]	PSNR [dB]
Frame (decoder)	**362.55**	**31.59**	**499.6**	**33.18**	**673.73**	**34.57**	**1076.73**	**37.61**
Block (decoder)	**362.32**	**31.59**	**499.37**	**33.18**	**664.32**	**34.58**	**1045.2**	**37.61**
Pixel (decoder)	**361.61**	**31.59**	**496.78**	**33.18**	**663.61**	**34.57**	**1049.44**	**37.61**
Frame (offline)	362.55	31.59	498.19	33.18	664.32	34.57	1042.38	37.6
Block (offline)	358.55	31.59	486.9	33.18	644.56	34.58	1015.09	37.61

As observed in Table 2, adapting dynamically, at the decoder, the Laplacian distribution parameter at the block level allows achieving better RD performance than the one obtained at the coarser frame granularity level. At the pixel level, the spatial region of support is only a pixel which can cause some instability; however the RD results are encouraging, especially for the 2nd bitplane where it outperforms the block and frame level estimation. In terms of computational complexity, the finest granularity level is the most demanding one; however, this should not be a critical issue since the complexity increase occurs at the decoder side where complexity is not a burden according to the WZ coding paradigm.

The online decoder α estimation algorithms proposed in this paper have a small loss in coding efficiency in comparison with the equivalent offline (encoder) methods presented in Section 3, both at the frame and block levels. This coding efficiency loss can be explained by noting that the offline process has access to the original frames while the online process only has an approximation of the true variance given the motion compensated residual. It is important to note however that the methods proposed here have the major advantage of being performed at the decoder leading to a more realistic WZ video coding scenario.

5. FINAL REMARKS

The techniques proposed in this paper alleviate the encoder from the high computational and cumbersome task of recreate the side information and allow the decoder to perform the estimation of the correlation noise distribution. These methods enable practical DVC solutions where the encoder has low complexity constraints. It is presented a complete analysis with both offline (encoder) and online (decoder) techniques, which work at different granularity levels, to estimate the correlation noise model. For future work is planned to combine the techniques proposed here with spatial coherence analysis of the side information frame to further enhance the RD efficiency.

6. REFERENCES

[1] ISO/IEC International Standard 14496-10:2003, "Information Technology – Coding of Audio-visual Objects – Part 10: Advanced Video Coding".

[2] J. Slepian, and J. Wolf, "Noiseless Coding of Correlated Information Sources", *IEEE Trans. on Inform. Theory*, Vol. 19, No. 4, July 1973.

[3] A. Wyner, and J. Ziv, "The Rate-Distortion Function for Source Coding with Side Information at the Decoder", *IEEE Trans. on Inform. Theory*, Vol. 22, No. 1, January 1976.

[4] A. Aaron, R. Zhang, and B. Girod, "Wyner-Ziv Coding for Motion Video", *Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, USA, November 2002.

[5] J. Ascenso, C. Brites and F. Pereira, "Motion Compensated Refinement for Low Complexity Pixel Based Distributed Video Coding", *IEEE International Conference on Advanced Video and Signal Based Surveillance*, Como, Italy, September 2005.

[6] J. Ascenso, C. Brites, and F. Pereira, "Improving Frame Interpolation with Spatial Motion Smoothing for Pixel Domain Distributed Video Coding", *5th EURASIP*, Slovak Republic, July 2005.

[7] A. Aaron, E. Setton and B. Girod, "Towards Practical Wyner-Ziv Coding of Video", *IEEE International Conference on Image Processing*, Barcelona, Spain, September 2003.