Modeling Correlation Noise Statistics at Decoder for Pixel Based Wyner-Ziv Video Coding *

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

[1]Instituto Superior Técnico – Instituto de Telecomunicações, Lisbon, Portugal
[2]Instituto Superior de Engenharia de Lisboa – Instituto de Telecomunicações, Lisbon, Portugal

Abstract. Distributed video coding (DVC) is a new video coding paradigm based on two key Information Theory results: the Slepian-Wolf and Wyner-Ziv theorems. Recently, promising results were shown in Wyner-Ziv (WZ) video coding, a particular case of DVC. In the literature, many practical WZ coding approaches model the correlation noise between the original frame and the so-called side information by a given distribution whose relevant parameters are estimated offline, at the encoder. This paper proposes an algorithm to estimate, at the decoder, and at the frame level, the correlation or error distribution between the original and the side information frames, in a way which is as efficient as the estimation made at the encoder based on the original information. This approach relieves the encoder from the task to perform this estimation based on the original information, which is rather important since DVC solutions are typically adopted under low encoder complexity constraints.

Keywords – Wyner-Ziv coding, distributed video coding, correlation noise model, frame level

1. INTRODUCTION

Today's digital video coding paradigm, represented by the standardization efforts of ITU-T VCEG and ISO/IEC MPEG, lies on hybrid DCT and interframe predictive coding. In this coding framework, the encoder is typically 5 to 10 times more complex than the decoder, mainly due to the motion estimation/compensation task. After all, it is the encoder that has to take all coding decisions, and is responsible to achieve the best rate-distortion (RD) performance, while the decoder remains a pure executer of the encoder "orders". This kind of architecture is well-suited for applications where the video is encoded once and decoded many times, i.e. one-to-many topologies, such as broadcasting or video-on-demand, where the cost of the decoder is more critical than the cost of the encoder.

* The work presented was developed within VISNET, a Network of Excellence (http://www.visnet-noe.org), and DISCOVER, a Future Emerging Technology project (http://www.discoverdvc.org/) both funded by the European Commission.

In recent years, with emerging applications such as wireless low-power surveillance, multimedia sensor networks, wireless PC cameras and mobile camera phones, the traditional video coding architecture is being challenged. These applications have very different requirements than those of the broadcast video delivery systems. For some applications, a low power consumption both at the encoder and decoder sides is essential, e.g. in mobile camera phones. In other types of applications, notably when there is a high number of encoders and only one decoder, e.g. surveillance, low cost encoder devices are necessary.

Distributed video coding, a new video coding paradigm, fits well in these scenarios, since it enables to explore the video statistics, partially or totally, at the decoder; thus, DVC enables a flexible allocation of the complexity burden between the encoder and the decoder. From the Information Theory, the Slepian-Wolf theorem [1] 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 (WZ) coding [2]. WZ coding deals with the lossy source coding of an X sequence considering that a dependent sequence Y, known as side information, is only available at the decoder. The side information is usually interpreted as an attempt of the decoder to obtain an estimate of the original frame.

One of the most interesting practical WZ approaches is the turbo-based pixel domain Wyner-Ziv coding scheme presented in [3], where all the source statistics are exploited at the decoder. In this solution, the decoder is the responsible to achieve compression following the Wyner-Ziv coding paradigm. The coding efficiency of Wyner-Ziv coding approaches depends critically on the capability to model the correlation noise between the original data and the side information generated at the decoder. Since the side information quality varies along time and the decoder does not have access to the original data, correlation noise statistics modeling at the decoder becomes a complex operation.

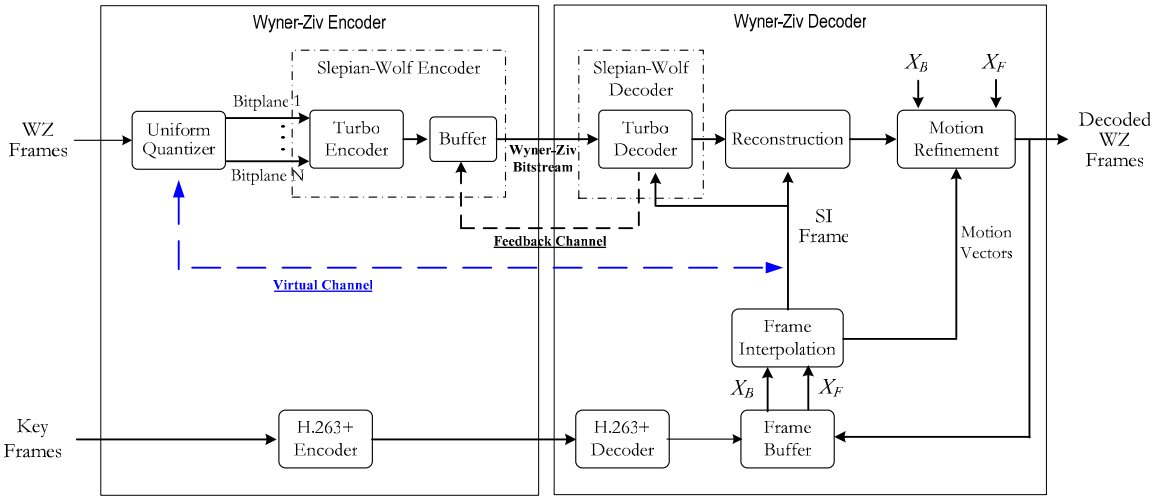

Fig. 1. IST-PDWZ video codec architecture.

In this paper, a new algorithm to estimate the correlation noise model based on temporal correlation information, in this case the motion compensated residual at the decoder, is proposed. This algorithm adaptively models the correlation noise distribution, at the decoder, at the frame level.

This paper is organized as follows: Section 2 presents a brief summary of the IST-PDWZ codec. In Section 3, a new approach to model the correlation noise statistics is described. Several experiments are performed, in Section 4, to evaluate and compare the coding efficiency of the proposed approach and, finally, in Section 5, conclusions and some future work topics are presented.

2. The IST-Pixel Domain Wyner-Ziv Video (IST-PDWZ) Codec

Figure 1 illustrates the architecture of the IST-PDWZ video codec proposed in [4]. Although this codec is based on the pixel domain Wyner-Ziv coding scheme proposed in [3], it includes major improvements, notably a more efficient side information (SI) creation solution at the decoder by using motion compensated frame interpolation with spatial motion smoothing (for more details consult [5]); the more accurate the side information is, the fewer are the Wyner-Ziv bits required to provide a reliable decoding of the Wyner-Ziv frame.

In a nutshell, the overall coding process is as follows: the video frames are organized into key frames and Wyner-Ziv frames. The key frames are encoded with a conventional intraframe codec with a quality similar to the quality of the WZ frames. Wyner-Ziv frames are encoded pixel by pixel; the pixels are quantized using a 2^M-level uniform scalar quantizer, generating the quantized symbol stream. Over the resulting quantized symbol stream bitplane extraction is performed and each bitplane is then independently turbo encoded. The turbo encoder encloses two recursive systematic convolutional (RSC) encoders of rate ½ and a pseudo-random interleaver. Each RSC encoder outputs the parity stream and the systematic stream. After turbo encoding a bitplane, the systematic part is discarded 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 SI (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. In this paper, longer GOP lengths, notably 4 and 8, are also considered; for these two GOP lengths, the SI frame is generated using both previously decoded key frames and WZ frames, according to the frame interpolation structure illustrated in Figure 2 for a GOP length of 4 [6]. Thus, the SI frame corresponding to the WZ frame WZ_2 is interpolated from the key frames K_0 and K_4, and so on. A similar frame interpolation structure is used for longer GOP lengths.

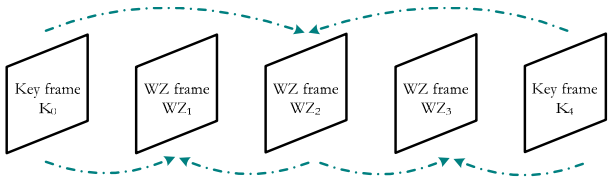

Fig. 2. Frame interpolation structure for GOP length of 4.

The side information 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. The refinement is performed with the help of the frames used to generate the side information and the motion vectors obtained by frame interpolation. As much as the authors know, this is among the best performing pixel domain WZ codec described in the literature, and was fully developed by the authors (including the turbo codec).

3. Proposing a Decoder-Generated Correlation Noise Statistics Model

In order to make good use of the side information (generated at the decoder by frame interpolation) in terms of coding efficiency, the decoder, notably the turbo decoder, needs to have reliable information on 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 since the side information may be seen as a 'corrupted' version of the original information.

In previous works, e.g. [4], [5], the authors used a Laplacian distribution as in (1) to model the statistical correlation between the original frame and the side information; this Laplacian distribution is used to convert the side information (pixel values) into soft-input information needed for turbo decoding.

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

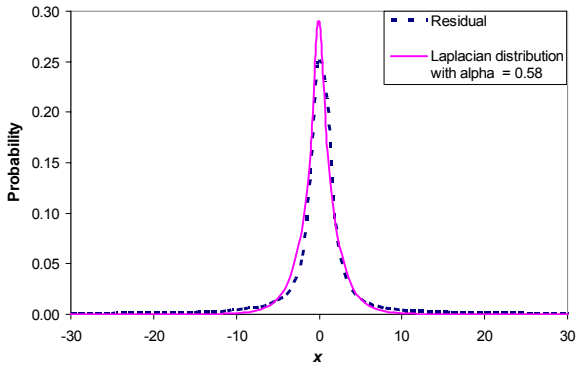

Fig. 3. Residual distribution for the *Foreman* QCIF video sequence.

Figure 3 depicts the actual distribution of the residual (WZ–SI), i.e. the luminance difference between corresponding pixels in the WZ frame and the side information frame for the *Foreman* QCIF video sequence. A Laplacian distribution given by (1) is also plotted in Figure 3, with the parameter α equal to 0.58. As it can be noticed, the Laplacian distribution is a good approximation of the residual (*WZ – SI*) distribution

In previous works, the Laplacian distribution parameter α, given by

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

has been computed offline, at the encoder, i.e. before the WZ coding procedure starts, over the whole video sequence and kept constant for the decoding of all WZ frames, after transmission to the decoder.

In (2), the parameter σ2 represents the variance between the original WZ frame and the SI frame. This process is not acceptable and efficient because it requires the encoder to recreate the side information (while WZ encoders should be of low complexity), and does not exploit the variability of the correlation model along time.

The main novelty of this paper resides in the dynamic variation of the Laplacian distribution parameter α at the frame level, after estimation at the decoder; the α value at the frame level is then used to obtain the soft-input information needed by the turbo decoder. The proposed α estimation algorithm is performed at the decoder, where more computational resources are available according to the WZ coding paradigm. Moreover these parameters do not have to be transmitted from the encoder to the decoder, typically under error prone conditions. This is an important departure from previous work in the literature which leads to a more practical Wyner-Ziv

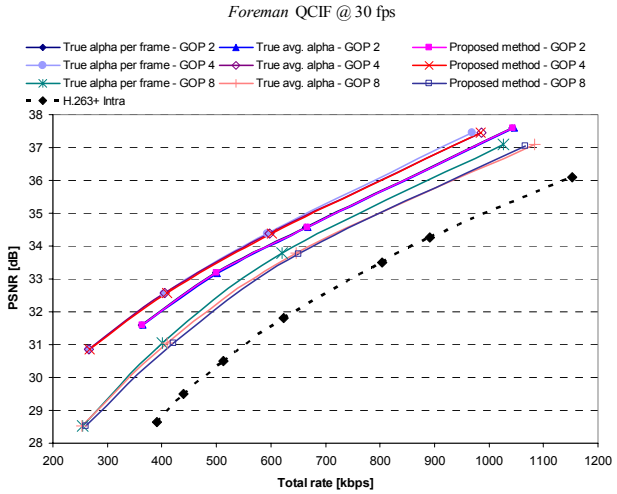

Fig. 4. IST-PDWZ RD performance for the *Foreman* QCIF sequence considering GOP lengths of 2, 4 and 8.

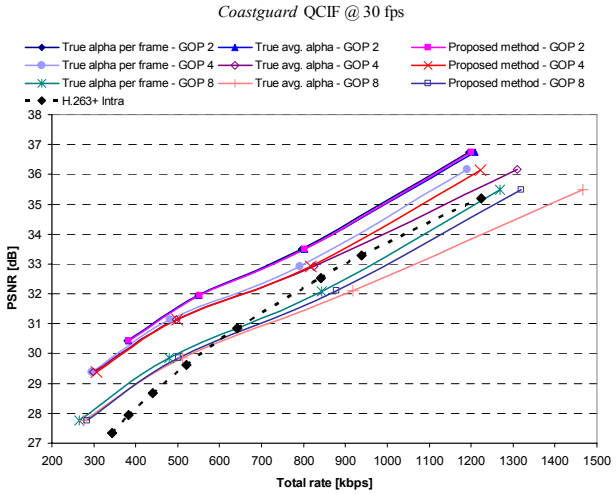

Fig. 5. IST-PDWZ RD performance for the *Coastguard* QCIF sequence considering GOP lengths of 2, 4 and 8.

video coding solution since it is no more necessary to recreate the side information at the encoder side.

The novel estimation approach proposed here makes use of X_B and X_F (where X_B and X_F can be previously decoded key frames or WZ frames, according to the GOP length – see Section 2) to obtain the α parameter estimate. Assuming linear motion between X_B and X_F, and that the interpolated frame is temporally equally spaced from X_B and X_F, the pixel values of both frames contribute with a ½ weight to the SI interpolated pixels. SI is thus given by (3), where 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 SI frame; (dx_b, dy_b) and (dx_f, dy_f) represent the motion vectors for X_B and X_F, respectively.

$$SI(x,y) = \frac{1}{2} X_B(x + dx_b, y + dy_b) + \frac{1}{2} X_F(x + dx_f, y + dy_f) \quad (3)$$

In order to estimate the Laplacian distribution parameter α at the decoder, it is necessary to define a variable which somehow expresses the variance σ^2 between the original and the side information, since the original information is not available at the decoder.

By computing an weighted mean squared error (WMSE) between X_B and X_F motion compensated, as described in (4), a confidence measure in the SI generation procedure is obtained; this confidence measure indicates how good the frame interpolation outcome is, i.e. how close the side information is to the corresponding original frame. When the motion compensated residual between X_B and X_F is high, it means that the interpolated frame presents a significant amount of errors when compared to the original frame, and thus a wide variance σ^2 should be considered in (2); on the other hand, small motion compensated residuals means a successful frame interpolation and a narrow variance is expected in (2).

In (4), the weight ½ indicates that both X_B and X_F frames contribute with a ½ weight to the SI interpolated values and L stands for the frame size (L = N×M). The WMSE metric given by (4) can therefore be used to represent the variance σ^2 between the original information and the side information.

$$WMSE(x,y) = \left(\frac{1}{2}\right)^2 \frac{\sum_{(x,y)\in SI} \left(X_B(x + dx_b, y + dy_b) - X_F(x + dx_f, y + dy_f)\right)^2}{L} \quad (4)$$

The α estimate for each WZ frame is thus obtained from (2) substituting σ^2 by the WMSE value calculated from (4).

4. Experimental Results

Figure 4 and Figure 5 show the IST-PDWZ RD results for the first 101 frames of the *Foreman* and *Coastguard* QCIF video sequences at 30 frames per second (fps). GOP lengths of 2, 4 and 8 have been used. Three configurations are considered in terms of the statistical modeling of the correlation noise:

i) True average α: this is the average in time of the α parameter computed offline, at the encoder, based on the residual histogram between the original

information and the corresponding side information; this α value is kept constant for the decoding of all WZ frames;

ii) True α: As above but now an α parameter is used for each WZ frame (it is called true α value because original WZ frames are used);

iii) Decoder α: The α parameter is obtained by the estimation algorithm proposed in this paper with the advantages already mentioned.

The test conditions for the frame interpolation and motion refinement modules are [4]:

- **Frame interpolation:** 8×8 block size, ±8 pixels for the search range of the full block motion estimation, and ±2 pixels for the search range of the bi-directional motion estimation.
- **Motion refinement:** Block size remains unchanged; the threshold to determinate if the block is motion compensated or not is an average difference of 0.15 per pixel and the search range is ±4 pixels.

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 frames).

As can be noticed in Figure 4 and Figure 5, the true α allows achieving better RD performance when compared to the performance of the true average α, mainly for GOP lengths longer than 2. Notice that the temporal spacing between the frames X_B and X_F used to interpolate the SI frame varies according to the position of the SI frame within the GOP (see Figure 2). This implies the SI frames within a GOP will have more or less interpolation errors depending on the temporal spacing between X_B and X_F; a large temporal spacing means that the frame interpolation fails more often and thus the quality of the side information decreases. This variation of the SI frame quality within the GOP explains why performing a more dynamic adaptation of the α value brings a better RD performance.

The α estimation algorithm proposed here presents some coding efficiency loss regarding the true α, mostly for GOP lengths longer than 2. The loss in the RD performance is mainly due to the fact that the original frames are not available at the decoder. The coding efficiency loss increases for longer GOPs, since some SI frames will accumulate motion interpolation errors (cases where X_B and X_F also result from frame interpolation and thus have interpolation errors associated to them) and this is not well modeled by the decoder α estimation algorithm. However the method proposed here corresponds to a more realistic WZ video coding scenario.

As shown in Figure 4, the IST-PDWZ RD performance is above the H.263+ intra coding for the first 101 frames of the *Foreman* sequence; this conclusion is independent of the configurations used in terms of the statistical modeling of the correlation noise. For long GOP lengths (e.g. 8), and for the *Coastguard* sequence, the IST-PDWZ codec presents a penalty in the coding efficiency when compared to the H.263+ Intra (Figure 5); this is mainly due to the high motion (i.e. the pan-left) that occurs in the *Coastguard* sequence. When high motion occurs in a sequence and long GOP lengths are used, the motion interpolation task is quite difficult. In this case, a small search range of ±8 pixels is not sufficient to capture such amount of motion and thus a high search range should be used. However, using a high search range, the frame interpolation fails more often when the amount of motion is low. Thus, a trade-off between the search range and the amount of motion to be captured must be defined.

5. Final Remarks

The main contribution of this paper is to present a novel, simple algorithm to estimate at the decoder, and at frame level, the error distribution between the SI and the original frame for the IST-PDWZ codec. The results obtained with this simple algorithm show a RD performance close to the one obtained with the true α value per frame computed offline using the original WZ data. As future work, it is planned to further enhance the RD performance by combining temporal correlation information with spatial coherence analysis of the side information frame, at different granularity levels, i.e. block and pixel. The combination of temporal and spatial information would enable to achieve a more accurate estimate of the correlation noise statistics model.

References

1. Slepian, J., Wolf, J.: "Noiseless Coding of Correlated Information Sources", IEEE Transactions on Information Theory, Vol. 19, No. 4, July 1973.
2. Wyner, A., Ziv, J.: "The Rate-Distortion Function for Source Coding with Side Information at the Decoder", IEEE Transactions on Information Theory, Vol. 22, No. 1, January 1976.
3. Aaron, A., Zhang, R., Girod, B.; "Wyner-Ziv Coding for Motion Video", Asilomar Conference on Signals, Systems and Computers, Pacific Grove,

USA, November 2002.

4. Ascenso, J., Brites, C., Pereira, F.: “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.
5. Ascenso, J., Brites, C., Pereira, F.: “Improving Frame Interpolation with Spatial Motion Smoothing for Pixel Domain Distributed Video Coding”, 5th EURASIP Conference on Speech and Image Processing, Multimedia Communications and Services, Slovak Republic, July 2005.
6. Aaron, A., Setton, E., Girod, B.: “Towards Practical Wyner-Ziv Coding of Video”, IEEE International Conference on Image Processing, Barcelona, Spain, September 2003.