Improved Correlation Noise Statistics Modeling in Frame-based Pixel Domain Wyner-Ziv Video Coding *

Alan Trapanese Marco Tagliasacchi Stefano Tubaro
Politecnico di Milano - Italy

João Ascenso Catarina Brites Fernando Pereira
Instituto Superior Técnico - Instituto de Telecomunicações, Lisbon - Portugal

Abstract

Distributed source coding principles have been recently applied to video coding in order to achieve a flexible distribution of the complexity burden between the encoder and the decoder. In this paper we elaborate on a pixel based Wyner-Ziv video codec that shifts all the complexity of the motion estimation phase to the decoder, thus achieving light encoding. In the literature, the statistics of correlation noise between the frame to be encoded and the motion-compensated side information available at the decoder is modeled as a Laplacian distribution. In this paper we elaborate on this topic and we show that a better model can be fitted, achieving a substantial coding efficiency gain. Moreover we discuss the effect of using a side information computed either from perfectly reconstructed (lossless) or from quantized neighboring frames.

1 Introduction

Today's video coding architectures are based on the "downlink" broadcast model, where the video content is encoded once and decoded multiple times. All the ITU-T VCEG and ISO/IEC MPEG standards follow this approach relying on the hybrid block-based motion compensation/DCT transform (MC/DCT) architecture. 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. 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. For example, 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, a new video coding paradigm, fits well in these scenarios, since it enables to explore the video statistics, partially or totally, at the decoder. Distributed video coding lays its foundations on distributed source coding principles stated by the Slepian-Wolf [1] and the Wyner-Ziv [2] theorems. Despite the theory has been well understood since the 70's, only recently practical video coding schemes have been presented targeting different application requirements ranging from low-encoding complexity, robustness to channel losses and scalability.

2 IST-PDWZ video codec architecture

The IST Pixel Domain Wyner-Ziv (IST-PDWZ) video codec we use in this paper [3] is based on the pixel domain Wyner-Ziv coding architecture proposed in [4]. However, there are major differences in the frame interpolation tools further discusses in [3]. This approach offers a pixel domain intra-frame encoder and inter-frame decoder with very low computational encoder complexity. When compared to traditional video coding, the proposed encoding scheme is less complex by several degrees of magnitude. Figure 1 illustrates the global architecture of the IST-PDWZ codec. In this architecture each even frame X_{2i} of the video sequence is called Wyner-Ziv frame and the two adjacent odd frames X_{2i-1} and X_{2i+1} are referred as key frames; in the literature [4] it is assumed that they are perfectly reconstructed (lossless) at the decoder. Each pixel in the Wyner-Ziv frame is uniformly quantized. Bitplane extraction is performed from the entire image and then each bitplane is fed into a turbo encoder. At the decoder, the motion-compensated frame interpolation module generates the side information, Y_{2i} [3], which will be used by the turbo decoder and reconstruction modules. The decoder operates in a bitplane by bitplane basis and starts by decoding the most significant bitplane and it only proceeds to the next bitplane after each bitplane is successfully turbo decoded (i.e. when most of

*The authors wish to acknowledge the support provided by the European Network of Excellence VISNET (http://www.visnetnoe.org)

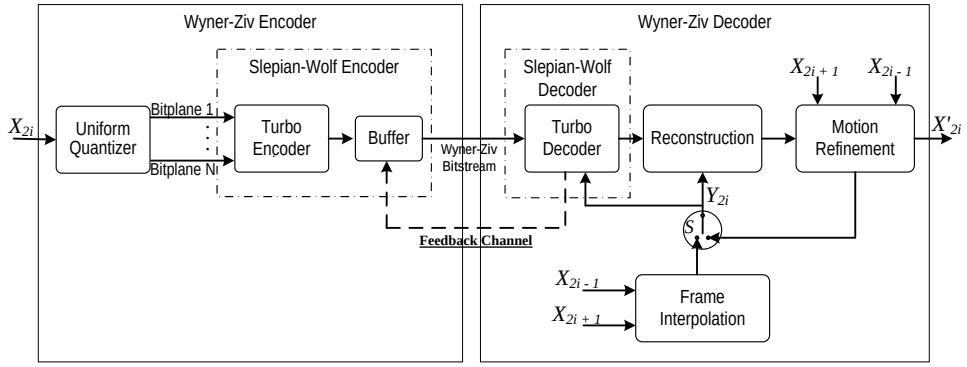

Figure 1: Block diagram of the IST-PDWZ codec

the errors are corrected).

2.1 Turbo codec overview

Let us define X_{2i}^j as the j^th bitplane of the Wyner-Ziv frame read in raster scan order and Y_{2i}^j the corresponding bitplane of the motion-intepolated side information. We can interpret X_{2i}^j as a binary codeword of length $rows \times cols$ and Y_{2i}^j is its noisy version. The IST-PDWZ architecture adopted in this paper uses a Slepian-Wolf rate compatible punctured turbo (RCPT) coder in order to correct the mismatch (errors) between the side information Y_{2i}^j and the source to be decoded X_{2i}^j.

As shown in Figure 1, the Slepian-Wolf encoder includes a turbo encoder and a buffer and it produces a sequence of parity bits (redundant bits) associated to each bitplane X_{2i}^j. In this architecture, two identical recursive encoders of rate $\frac{1}{2}$ are used; this means that for each information bit, two parity bits are produced. The parity bits generated by the turbo encoder are then stored in the buffer, punctured and transmitted upon request by the decoder while the systematic bits (X_{2i}^j) are discarded. The puncturing operation allows sending only a fraction of the parity bits and follows a specific puncturing pattern. The feedback channel is necessary to adapt to the changing statistics between the side information and the frame to be decoded, i.e. to the quality (or accuracy) of the frame interpolation or motion refinement process.

At the decoder, the iterative MAP (Maximum A Posteriori) turbo decoder employs a Laplacian noise model to help the error correction capability of the turbo codes. An ideal error detection capability is also assumed at the decoder, i.e. the decoder is able to measure in a perfect way the current bitplane error rate, Pe. In the following section we elaborate on the modeling of the statistical distribution of the correlation noise, showing that a better understanding of the model comes with improved rate-distortion performance.

3 Improved correlation noise model

In the literature [4][3] the key frames used in input to the motion interpolation are assumed to be perfectly known at the decoder (lossless coding). This hypothesis is rather unpractical in real applications for two reasons: the reconstructed sequence exhibits quality fluctuations; the key frames would require a bitrate budget much larger than Wyner-Ziv frames, as the former are lossless encoded. In this paper we depart from this scenario investigating what is the effect of computing the side information from quantized key frames, which is a much more realistic situation.

In our experiments we choose for the key frame a quantization step size that gives approximately the same quality as the Wyner-Ziv frames for a given number of decoded bitplanes. We found that a QP ($qstep = 2 \cdot QP$ [1]) equal to 13, 10, 8, 5 work quite well when we decode respectively up to the 1^{st}, 2^{nd}, 3^{rd} and 4^{th} most significant bitplane. Figure 4 shows that there is a significant coding efficiency drop when the side information is quantized (WZ lossless key frames vs WZ lossy key frames). This result is reasonable if we compare the quality of the motion-compensated side information used as a starting point by the turbo decoder. Figure 2 shows the PSNR of the side information for each frame of the *Foreman* and *Coastguard* sequences, when the lossless or lossy key frames are used. Table 1 reports the averages over the whole sequences.

The model used to describe the statistics of the correlation noise between the Wyner-Ziv frame and the motion-interpolated side information needs to be adjusted. First of all we need to take into consideration the effect of quantization noise. When the side information is perfectly reconstructed at the decoder the correlation noise is simply $N_{2i} = X_{2i} - Y_{2i}$. By considering quantization noise, the

[1] as in H.263+ standard

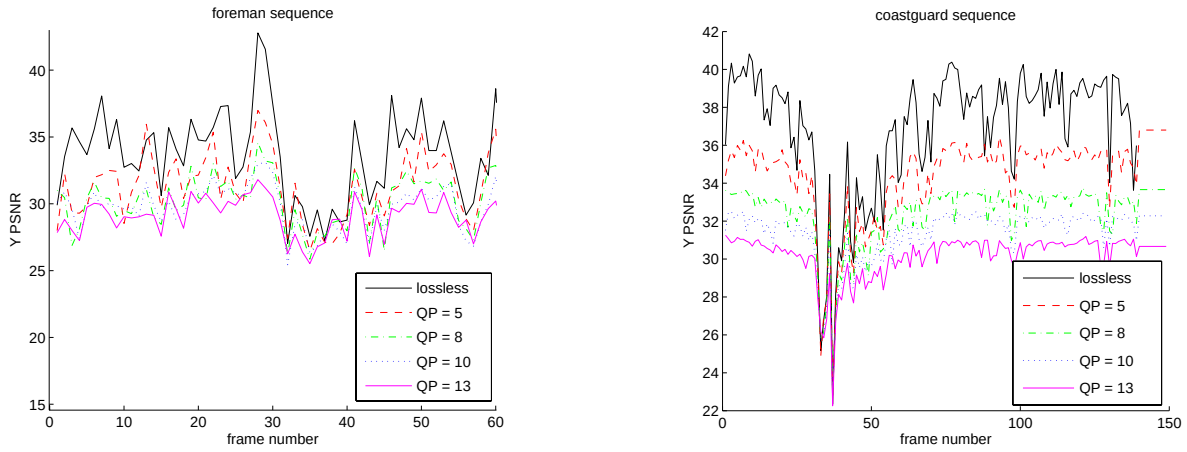

Figure 2: Quality of the motion-compensated interpolation: lossless vs lossy. The frame index refers to the interpolated frames only (odd frames). Left: *Foreman* sequence. QCIF@30fps (only the first 60 odd frames of *Foreman* are shown to avoid cluttering the figure). Right *Coastguard* sequence. QCIF@30fps

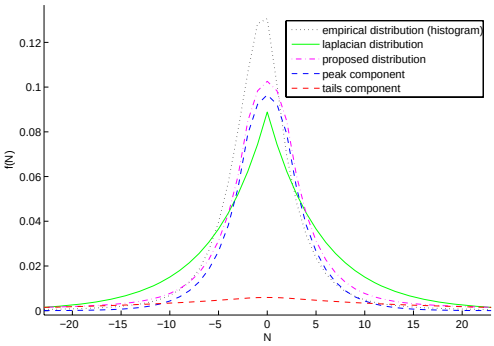

Figure 3: *Foreman* sequence. Statistical model of the correlation noise

previous expression turns out to be:

$$\begin{aligned}\hat{N}_{2i} = X_{2i} - \hat{Y}_{2i} &=^{(a)} X_{2i} - \frac{\hat{Y}_{2i-1} + \hat{Y}_{2i+1}}{2} = \\ &=^{(b)} X_{2i} - \frac{Y_{2i-1} + Q_{2i+1} + Y_{2i+1} + Q_{2i-1}}{2} = \\ &= X_{2i} - Y_{2i} - \frac{Q_{2i+1} + Q_{2i-1}}{2} = \\ &= N_{2i} - \frac{Q_{2i+1} + Q_{2i-1}}{2},\end{aligned}$$

where (a) comes from the fact that the side information $\hat{Y}_{2i}$ is computed by motion-compensated interpolation of the quantized key frames $\hat{Y}_{2i-1}$ and $\hat{Y}_{2i+1}$ and in (b) Q_{2i+1} and Q_{2i-1} are the quantization noise terms associated with the key frames. Here we are assuming that, at high rates, the quantization noise is uncorrelated with the source and that

Table 1: *Y PSNR of the motion-compensated interpolation averaged over the whole sequence*

sequence	lossless	QP = 5	QP = 8	QP = 10	QP = 15
Foreman	33.6	31.2	30.1	29.6	28.7
Coastguard	36.9	34.2	32.3	31.3	30.1

the two terms are independent from each other and from the correlation noise N_{2i}. If we assume Q_{2i+1} and Q_{2i-1} to be independent and have uniform distribution, the statistical distribution of $\hat{N}$ is $f_{\hat{N}} = 0.5 \cdot f_N * f_{Q_{2i+1}} * f_{Q_{2i-1}}$, where $*$ denotes the convolution product.

We found out that the Laplacian model used in [4][3] to represent the distribution of N_{2i} is not the best choice, as the tails of the model go to zero slower that the empirical distribution (see Figure 3). Correct modeling of the tails turns out to be crucial to help the turbo decoding process. In fact when the tails vanish too slowly, the turbo decoder tends to assign a higher likelihood to values that are far apart from the corresponding side information, increasing the chance to decode outliers. We tried to fit a generalized Gaussian distribution without getting significant results though. On the other hand, we found that the following model works quite well in practice:

$$f_N(n) = K \cdot \left[k_1 \frac{\alpha_1}{2} e^{-\alpha_1 |n|} + k_2 \frac{\alpha_2}{2} e^{-\alpha_2 |n|} \right]$$

We could not find a closed form expression of the unknown parameters using a maximum likelihood estimation approach. Nevertheless we empirically set $k_1 = 1$,

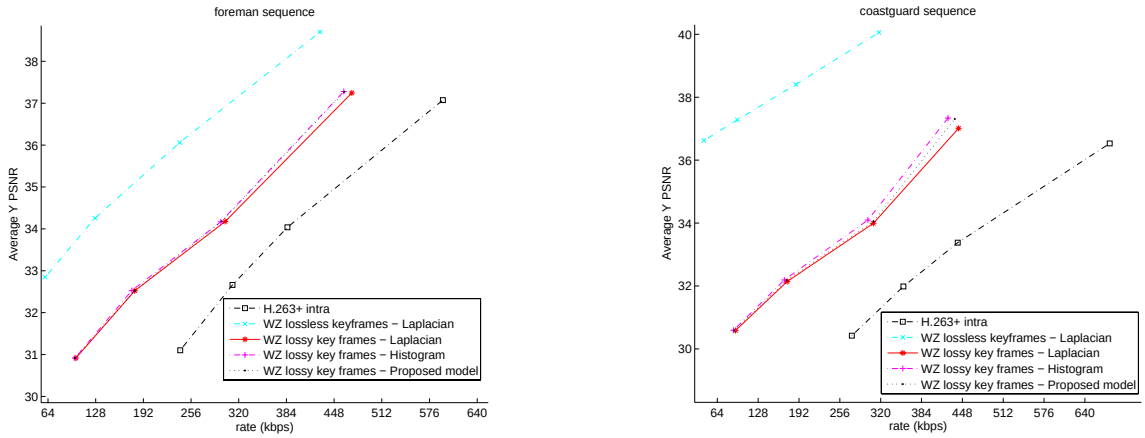

Figure 4: Left: *Foreman* sequence. QCIF@30fps. 400 frames. Right *Coastguard* sequence. QCIF@30fps. 300 frames

$k_2 = 1/4$, and:

$$\alpha_1 = \sqrt{2}/\left[\frac{1}{M}\sum_{i=0}^{M-1}|n_i|\right] \quad \alpha_2 = \sqrt{2}/\left[\sqrt[4]{\frac{1}{M}\sum_{i=0}^{M-1}|n_i|^4}\right]$$

Since in general $\alpha_1 > \alpha_2$ (by the Jensen inequality), the first term is used to correctly model the peak of the empirical distribution, whereas the second to raise the tails.

Figure 3 compares the statistical distribution of the correlation noise used in [4][3] with the actual sample distribution (histogram) and the proposed model for the *Foreman* sequence. The figure refers to the case when four bitplanes are decoded ($QP = 5$).

4 Experimental results

We carried out extensive experimental results on the *Foreman* and *Coastguard* sequences in order to evaluate the effect of the new statistical model of the correlation noise. Figure 4 shows that a coding gain of up to 0.4dB on average is observed for *Foreman* and 0.5dB for *Coastguard* due to the new statistical model. Note that with the proposed model we are only 0.2dB off the coding gain that can be obtained using the empirical distribution (histogram) of the correlation noise. The gain observed on the single frames is higher (up to 1.5dB) whenever the motion-compensated interpolation fails to correctly reconstruct the WZ-frame. We argue that this is due to the fact that the proposed model provides a better representation of the tails of the distributions.

5 Conclusions

In this paper we describe a model that better matches the statistical distribution of the correlation noise between the frame to be decoded and the motion-compensated side information, resulting in an improved rate-distortion performance. Our future work will focus on adapting the statistical model both in the spatial and in the temporal domain.

References

[1] J. D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, pp. 471–480, July 1973.

[2] A. D. Wyner and J. 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] J. Ascenso, C. Brites, and F. 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.

[4] A. Aaron, R. Zhang, and B. Girod, "Wyner-Ziv coding of motion video," in *Proceedings of the* 36^{th} *Asilomar Conference on Signals, Systems, and Computers*, vol. 1, (Pacific Grove, CA), pp. 240–244, October 2002.