EXPLOITING SPATIAL REDUNDANCY IN PIXEL DOMAIN WYNER-ZIV VIDEO CODING*

M. Tagliasacchi, A. Trapanese, S. Tubaro

Dipartimento di Elettronica e Informazione
Politecnico di Milano,
Milan - Italy

J. Ascenso, C. Brites, F. Pereira

Instituto Superior Técnico
Instituto de Telecomunicações,
Lisbon - Portugal

ABSTRACT

Distributed video coding is a recent paradigm that enables a flexible distribution of the computational complexity between the encoder and the decoder building on top of distributed source coding principles. In this paper we focus on the scenario where most of the complexity is shifted to the decoder, thus achieving light encoding. We elaborate on a well known pixel based Wyner-Ziv architecture and we improve its coding efficiency by exploiting both spatial and temporal correlation at the decoder side, without the need of performing any transform at the encoder. In order to generate the side information, the decoder adaptively chooses spatial or temporal information, based on the local estimate of the correlation noise. Simulations on test sequences demonstrate that a coding gain of up to +1.8dB can be obtained with respect to the case that generates the side information by motion interpolation only.

Index Terms— Video coding, motion analysis

1. INTRODUCTION

Today's digital video coding paradigm, represented by the ITU-T VCEG and ISO/IEC MPEG standardization efforts, relies on inter-frame predictive coding and block-based DCT transform in order to exploit both the temporal and spatial redundancy present in the video sequence. In this framework, the encoder has a higher computational complexity than the decoder (typically 5 to 10 times more complex). This is mainly due to the motion estimation and mode decision tools used to efficiently explore the temporal correlation. In fact, the encoder is responsible for all coding decisions to attain optimal rate-distortion (RD) performance, while the decoder remains a pure executer of the encoder "orders". This type 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. 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 different requirements than those of traditional video delivery systems. For some applications, it is essential to have low power consumption both at the encoder and at the decoder, e.g. in mobile camera phones. In other cases, notably when there are several encoders and only one decoder, e.g. in video surveillance applications, low complexity encoder devices are needed, possibly at the expense of a high-complexity decoder. While shifting the complexity burden from the encoder to the decoder, it is important to achieve a coding efficiency comparable with the state-of-the-art hybrid video coding schemes (e.g. the recent H.264/AVC standard [1]). This is currently rather far from being achieved and much research needs to happen in this area.

The main contribution of this paper is an algorithm that allows to exploit spatial redundancy in a Wyner-Ziv video coding architecture while working in the pixel domain, i.e. without recurring to transform based coding tools.

2. PDWZ VIDEO CODEC ARCHITECTURE

The Pixel Domain Wyner-Ziv (PDWZ) video codec we use in this paper is based on the pixel domain Wyner-Ziv coding architecture proposed in [2]. This coding architecture 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 PDWZ codec. 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 [2] it is assumed that they are perfectly reconstructed (lossless) at the decoder. In this paper as well as in our previous work [3][4] we consider a more realistic scenario by lossy encoding the key frames in such a way that the quality of the output sequence is kept constant. Each pixel in the Wyner-Ziv frame is uniformly quantized. Bitplane extraction is performed from the entire image and then each bit-plane is fed into a turbo encoder to generate a sequence of parity bits. At the decoder, the motion-compensated frame interpolation module generates the side information, Y_{2i} (see [5] for more details), which will be used by the turbo decoder and reconstruction modules. The decoder operates in a bit-plane by bit-plane basis and starts by decoding the most significant bit-plane and it only proceeds to the next bit-plane after each bit-plane is successfully turbo decoded (i.e. when most of the errors are corrected).

3. PROPOSED ALGORITHM

The coding architecture described in Figure 1 does not take advantage of spatial redundancy. In order to overcome this limitation, a transform domain Wyner-Ziv codec has been proposed in [6][7]. The Wyner-Ziv frame is DCT transformed, and transform coefficients of different blocks of the same order are grouped together to form frequency bands. Parity bits are generated separately for each frequency band and sent upon request to the decoder.

*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.

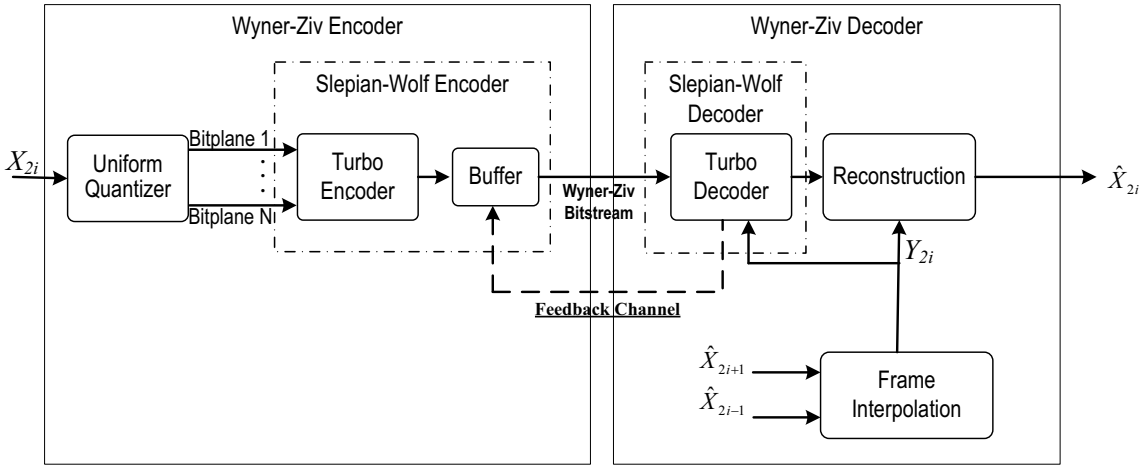

Fig. 1. PDWZ video codec architecture

In this paper we address spatial redundancy from a different angle, without introducing any transform at the encoder side. The goal here is to exploit spatial redundancy at the decoder side only, thus keeping the complexity of the encoding as low as possible.

Figure 2 illustrates a block diagram of the proposed algorithm. At the encoder, we split the Wyner-Ziv frame into two subsets X_{2i}^A and X_{2i}^B (the extension to a larger number of subsets is straightforward) based on a checkerboard pattern. Each subset is encoded independently as described in Section 2. At the decoder, subset X_{2i}^A is decoded first, using the side information obtained by motion interpolation Y_{2i}^T, thus exploiting only temporal correlation. Then, subset X_{2i}^B is decoded. In this case, pixel-by-pixel, the side information Y_{2i}^{ST} can be selectively chosen between the motion-interpolated frame Y_{2i}^T (to exploit temporal correlation) or by interpolating the previously decoded subset $\hat{X}_{2i}^A$ (to exploit spatial correlation).

The proposed algorithm can be detailed as follows:

1. *Frame splitting*: Let (x, y) be the coordinate values of a pixel:

$$\begin{aligned} &\text{If } [x \bmod 2]\text{xor}[(y+1) \bmod 2], \\ &\quad \text{then } (x,y) \in A, \\ &\quad \text{else } (x,y) \in B. \end{aligned} \tag{1}$$

   Denote with X_{2i}^A the pixel values assumed by X_{2i} in the pixel locations belonging to the set A.

2. *WZ encoding*: The encoder processes the two subsets independently, generating parity bits for both of them.

3. *Temporal side information generation*: The side information Y_{2i}^T is obtained by motion interpolation of $\hat{X}_{2i-1}$ and $\hat{X}_{2i+1}$ according to the algorithm described in [5].

4. *WZ decoding subset* X_{2i}^A, as described in Section 2 using Y_{2i}^T as side information.

5. *Spatial side information generation*: The spatial side information Y_{2i}^S is obtained by interpolating pixel values in $\hat{X}_{2i}^A$ in the pixel locations in B. A simple non-linear, adaptive algorithm is used to this purpose. With respect to Figure 3, the interpolation algorithm is the following:

   (a) Order the pixel values in the neighborhood $\mathcal{N}(x, y)$ of the current pixel. $\mathcal{N}(x, y) = \{(x-1, y), (x+1, y), (x, y-1), (x, y+1)\}$

   (b) $p(x, y)$ is set equal to the arithmetic mean of the two central values in the ordered list.

   Experimental results have shown that this scheme allows an improvement in the range 0.8-1.2 dB with respect to the case where a simple averaging of the four neighbors is performed.

6. *Spatio-temporal side information generation*: The spatial side information Y_{2i}^S is combined with the temporal side information Y_{2i}^T on a pixel-by-pixel basis in order to find the best side information. Ideally, the closest one to the original frame should be taken as the final side information, i.e.

$$Y_{2i}^{ST*}(x,y) = \arg \min_{j=S,T} |X_{2i}(x,y) - Y_{2i}^j(x,y)|^2 \tag{2}$$

   In practice, the decoder does not have access to X_{2i}. Therefore we need to infer $j = S, T$ for each (x, y) in the set B from the available information, i.e. $\hat{X}_{2i-1}$, $\hat{X}_{2i+1}$ and $\hat{X}_{2i}^A$. For each pixel (x, y) in B:

   (a) Compute the difference between Y_{2i}^T and $\hat{X}_{2i}^A$ as an estimate of the temporal correlation noise.

$$e(x,y) = \sum_{(m,n) \in A \cap \mathcal{N}(x,y)} |Y_{2i}^T(m,n) - \hat{X}_{2i}^A(m,n)|^2 \tag{3}$$

   (b) Generate the side information $Y_{2i}^{ST}(x, y)$:

$$\begin{aligned} &\text{If } e(x,y) < \epsilon, \quad Y_{2i}^{ST}(x,y) = Y_{2i}^T(x,y), \\ &\text{else } Y_{2i}^{ST}(x,y) = Y_{2i}^S(x,y) \end{aligned} \tag{4}$$

   where ϵ is a properly defined constant (we set $\epsilon = 128$ in our simulations) The idea is that temporal correlation is not spatially stationary, but its statistics slowly vary across space. Therefore we can use the observed temporal correlation for neighboring pixels as an estimate of the actual temporal correlation for the current pixel. When the estimated correlation is high, i.e. $e(x, y) < \epsilon$, the temporal side information Y_{2i}^T is used. Otherwise, the spatial side information Y_{2i}^S is used.

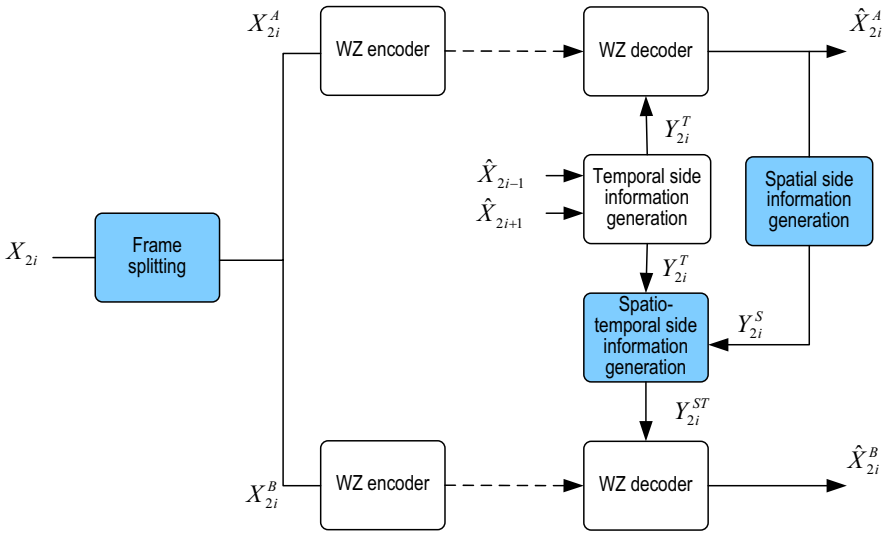

Fig. 2. Block diagram of the proposed coding algorithm.

7. *WZ decoding subset* X^B_{2i}, as described in Section 2 using Y^{ST}_{2i} as side information.

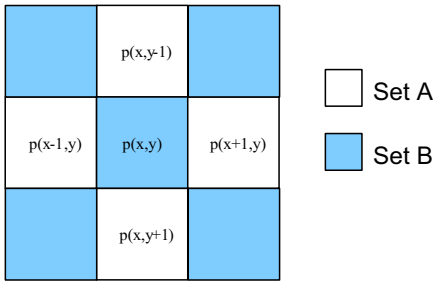

Fig. 3. Pixel neighborhood used for interpolation.

4. EXPERIMENTAL RESULTS

We carried out several experiments in order to showcase the promise of the proposed algorithm. First, we tested the quality of the side information that is used in the Wyner-Ziv decoding according to the coding architecture described in Section 2. The performance of the turbo decoding process heavily depends on the quality of the side information. Intuitively, a higher number of parity bits will be requested by the decoder when the correlation is weak, as more errors need to be corrected.

Table 1 indicates the quality of the side information measured in terms of PSNR (dB). In this experiment, both spatial and temporal interpolation is performed using original (not quantized) pixel values. A number of interesting conclusions can be drawn:

- The quality of the temporal side information Y^T_{2i} is always better than the spatial side information Y^S_{2i}.

	Y^T_{2i}	Y^S_{2i}	Y^{ST*}_{2i}	Y^{ST}_{2i}
Foreman	32.2	30.7	36.3	33.0
Coastguard	34.2	28.8	36.5	34.2
Mother&Daughter	38.1	35.7	40.6	38.3
News	33.0	28.0	38.2	34.5
Hall Monitor	36.7	30.1	39.3	36.8

Table 1. Comparison between the different types of side information used

- The proposed algorithm used to compute the spatio-temporal side information Y^{ST}_{2i} tends to improve the quality of the side information Y^T_{2i}. Nevertheless, the gain is sequence dependent. In sequences characterized by simple motion, i.e. *Coastguard*, *Mother&Daughter* and *Hall Monitor*, the gain is small (up to 0.2dB on average). When the complexity of the motion increases, temporal correlation usually decreases, thus spatial redundancy can be exploited. This justifies the gain of +0.8, +1.5 for *Foreman* and *News*.
- The estimate Y^{ST}_{2i} provided by the proposed algorithm is still far from the best side information that can be ideally computed by equation (2), denoted by Y^{ST*}_{2i}. The gap is pretty large for all sequences (up to 3.7dB for *News*), thus suggesting that more work can be done in this area by improving the criteria employed to switch between spatial and temporal side information.

Figure 4 shows the rate-distortion curves obtained integrating the proposed algorithm into the Wyner-Ziv video coding architecture. We notice that the improvements in the side information quality lead to a coding gain that depends on the sequence. The gain is up to +0.9dB for *News*, +0.6dB for *Foreman*, +0.6dB for *Mother&Daughter* (all at QCIF resolution, 30fps) when the spatio-temporal side information is used instead of only temporal side information. The gain is proportional to the improvement in the quality of the side information reported in Table 1. We recall that only half of the decoded

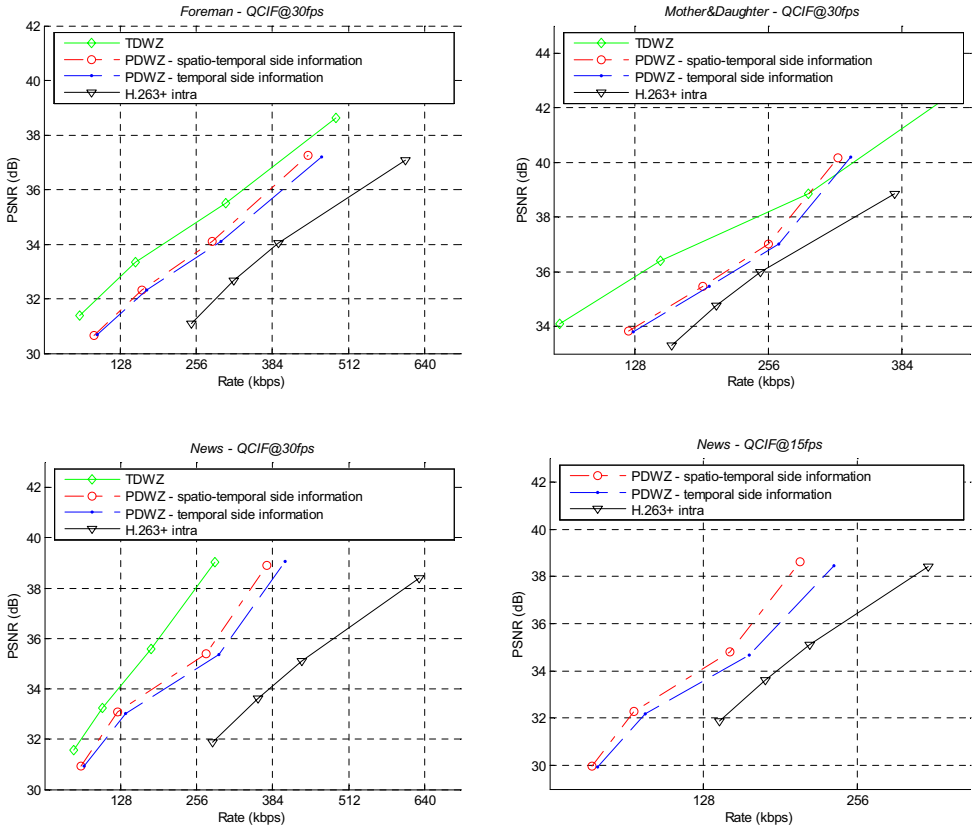

Fig. 4. Rate-distortions curves for the proposed codec.

pixels actually takes advantage of the proposed scheme, because the first half is encoded and decoded using only the temporal side information.

We have also carried out experiments with sequences at 15fps. In this case, inter-frame temporal correlation tends to be weaker and we expect larger gains from exploiting spatial redundancy. In fact, the gain is as large as +1.1dB for *Foreman* and +1.8dB for *News* (see Figure 4) with respect to the case that uses only the temporal side information Y_{2i}^{T}. Nevertheless, in this case the overall coding efficiency is decreased, and the coding gain over H.263+ intra is smaller.

In Figure 4 we also show the rate-distortion curve for a Transform-Domain Wyner-Ziv (TDWZ) codec [6][8]. The TDWZ codec almost always outperforms PDWZ codec, also when the spatio-temporal side information is used. Nevertheless, without adding complexity at the encoder, enhancing the side information with spatial data allows to partially bridge the gap between PDWZ and TDWZ.

5. CONCLUSIONS

This paper proposes a novel pixel domain algorithm to improve the quality of the side information in a Wyner-Ziv coding architecture. The algorithm leverages both spatial and temporal information. Ongoing research activities are currently focused on a generalization of the proposed algorithm, using more than two subsets, as well as evaluating the impact of this approach when the temporal distance between successive key frames is greater than two.

6. REFERENCES

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

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

[3] A. Trapanese, M. Tagliasacchi, S. Tubaro, J. Ascenso, C. Brites, and F. Pereira, "Embedding a block-based intra mode in frame-based pixel domain wyner-ziv video coding," in *Internationa Workshop on Very Low Bitrate Video Coding*, Costa del Rei, Sardinia, Italy, September 2005.

[4] A. Trapanese, M. Tagliasacchi, S. Tubaro, J. Ascenso, C. Brites, and F. Pereira, "Improved correlation noise statistics modeling in frame-based pixel domain wyner-ziv video coding," in *Internationa Workshop on Very Low Bitrate Video Coding*, Costa del Rei, Sardinia, Italy, September 2005.

[5] 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.

[6] A. Aaron, S. Rane, E. Setton, and B. Girod, "Transform-domain Wyner-Ziv codec for video," in *Visual Communications and Image Processing, Proc of SPIE*, San Jose, CA, January 2004.

[7] B. Girod, A. Aaron, S. Rane, and D. Monedero, "Distributed video coding," *Proceedings of the IEEE*, vol. 93, pp. 71–83, January 2005.

[8] C. Brites, J. Ascenso, and F. Pereira, "Improving transform domain wyner-ziv video coding performance," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, May 2006.