INTRA MODE DECISION BASED ON SPATIO-TEMPORAL CUES 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 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. We observe that the correlation noise statistics describing the relationship between the frame to be encoded and the side information available at the decoder is not spatially stationary. For this reason we introduce a mode decision scheme either at the encoder or at the decoder in such a way that when the estimated correlation is weak we opt for intra coding on a block-by-block basis. Both spatial and temporal criteria are used to determine whether a block is better intra coded or not.

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 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; this paper is contribution in that direction and it is based on our previous work appeared in [2][3][4].

*THE WORK PRESENTED WAS DEVELOPED WITHIN VISNET, A NETWORK OF EXCELLENCE (HTTP://WWW.VISNET-NOE.ORG), FUNDED BY THE EUROPEAN COMMISSION

2. IST-PDWZ VIDEO CODEC ARCHITECTURE

The IST Pixel Domain Wyner-Ziv (IST-PDWZ) video codec we use in this paper is based on the pixel domain Wyner-Ziv coding architecture proposed in [5]. However, there are major differences in the frame interpolation tools [2] and in the intra mode decision discussed in this paper. 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 IST-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 [5] 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 bitplane 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 [2] for more details), 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 errors are corrected). In Figure 1 the shaded blocks are responsible to perform the adaptive block based intra coding described in detail in the next section.

3. ADAPTIVE BLOCK BASED INTRA CODING IN WZ-FRAMES

Following [5], the turbo codec assumes that the correlation noise statistics between the source to be encoded X_{2i} and its side information Y_{2i} is spatially stationary neglecting any dependency on the spatial location. This simplification derives from the (false) assumption that the quality of the motion-interpolation estimate is constant across the whole frame. This section departs from the spatial stationarity assumption by acknowledging that covered/uncovered regions,

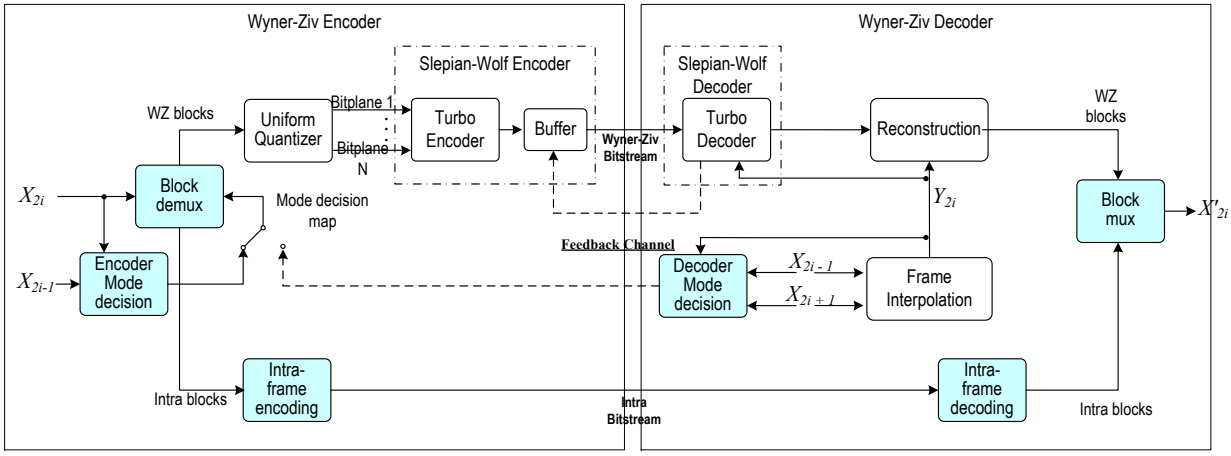

Fig. 1. IST-PDWZ video codec architecture

illumination changes and camera noise do affect significantly the quality of the motion interpolation phase. Although more complex coding schemes might try to adapt the correlation statistics locally on a block-by-block basis, i.e. providing an estimate of the correlation noise variance for each block, this paper considers a simpler approach. In this approach an intra-frame encoding mode is used for blocks whose correlation with the side information is weak; of course, this lack of correlation has to be detected without increasing significantly the encoder complexity. In the proposed solution, the encoding of a Wyner-Ziv frame proceeds as follows:

- Each WZ-frame is divided into non-overlapping 8×8 blocks
- For each block, a binary decision is made either at the encoder (see Section 3.1) or at the decoder (see Section 3.2); this decision indicates if the block will be Wyner-Ziv or intra coded.
  - For intra coded blocks, a DCT-based approach similar to H.263+ intra mode is used.
  - Wyner-Ziv blocks are encoded as explained in Section 2, where the only difference is the scanning order that skips intra encoded blocks

In the following sections, the criteria used to perform the mode decision either at the encoder or at the decoder will be described. The mode decision is performed for each WZ-frame, and a binary map, where each entry in the map indicates whether the block should be intra or Wyner-Ziv coded, is constructed. In order to efficiently code this information, a simple entropy coding algorithm is used. Since a strong spatial correlation between the coding mode of neighboring blocks is observed, a run-length encoding algorithm identifies runs of blocks assigned to the same mode in raster scan order and entropy encodes the symbol (mode, run-length) using a variable length code adapted to the statistics of the symbols. For QCIF sequences at 30fps, 396 binary decisions (one for each 8×8 block) need to be represented for each Wyner-Ziv frame, adding up to approximately 5.8kbps if transmitted uncoded. By run length encoding this information, it is possible to reduce the cost to 1-1.5kbps.

3.1. Intra Mode Decision at the Encoder

First, the case where the block-based Wyner-Ziv vs. intra decision is made at the encoder is considered. As a low encoding complexity scenario is targeted, the encoder is constrained not to perform any motion search. Therefore, a criterion that is easy to compute yet able to infer the quality of the motion-compensated side information available only at the decoder is needed. Two measures are taken into account in order to determine whether the block needs to be encoded in intra-frame mode.

1. The first metric computes the SAD (Sum of Absolute Differences) as an indication of the temporal coherence between the block B in the Wyner-Ziv frame and the co-located block in the previous key frame:

$$SAD_{enc} = \sum_{(x,y)\in B} |X_{2i}(x,y) - X_{2i-1}(x,y)| \quad (1)$$

2. The second metric measures the spatial smoothness of the block, by computing the pixel variance of the luminance component:

$$\sigma^2_{enc} = \frac{1}{M}\left[\sum_{(x,y)\in B} |X_{2i}(x,y)|^2 - \left(\sum_{(x,y)\in B} X_{2i}(x,y)\right)^2\right] \quad (2)$$

If $SAD_{enc} \geq T^{TC}_{enc}$ or if $\sigma^2_{enc} \leq T^{SS}_{enc}$, the block is intra coded. The rationale behind this decision is that when the zero-motion temporal correlation measured by SAD_{enc} is high, then it is likely that the correlation with the motion-interpolated side information is weak, therefore intra-frame coding is performed. On the other hand, when the block is smooth, thus the pixel variance is low, intra-frame encoding can efficiently tackle spatial redundancy. With this modification, the DVC coding scheme is no more strictly speaking an intra-frame encoder as it requires an extra buffer to store the previous key frame, needed to compute equation (1). The encoding process of WZ-frames proceeds as follows:

1. For each block, evaluate (1) and (2) and perform the mode decision (Wyner-Ziv or intra).

2. Encode and send the binary mode decision map to the decoder.
3. Encode the intra blocks as in H.263+ intra mode (DCT, scalar quantization, VLC entropy coding) and send the bits to the decoder.
4. Concatenate the pixels of all the WZ-blocks (each block is read in raster scan order before proceeding to the next block) and then encode the WZ-blocks by extracting the bitplanes to be Wyner-Ziv encoded.
5. Run the turbo encoder for each bitplane and store the parity bits in a buffer.

The decoding process turns out to be:

1. Build the side information Y_{2i} starting from the neighboring key frames X_{2i-1} and X_{2i+1}.
2. Receive and decode the mode decision map.
3. Receive and decode the intra coded blocks.
4. For each WZ-block, and for each bitplane, starting from the most significant one, correct the side information Y_{2i} to reconstruct X_{2i} by requesting parity bits and running the turbo decoder until the error probability is below the predefined threshold.

The intra vs. Wyner-Ziv mode decision at the encoder tends to make some incorrect choices, classifying in intra-frame coding mode some blocks also when the side information is a good estimate of the original frame, e.g. due to camera panning that cannot be captured without motion estimation. In order to overcome this limitation, a proposal to shift the mode decision to the decoder side is made, as detailed in the next section.

3.2. Intra Mode Decision at the Decoder

At the decoder, it is possible to better estimate the correlation statistics in sequences characterized by significant motion, as more complexity is allowed and thus motion estimation may be performed. The motion interpolation algorithm adopted in the IST-PDWZ codec assigns to each WZ-frame block a direct motion vector, i.e. the same motion vector is applied with reversed signs to indicate the predictor blocks in the previous and next key frames. Here the following metrics are proposed:

1. Although the decoder does not know the frame to be decoded, it can infer the quality of the motion interpolation by looking at the consistency between the forward and backward predictors, for example by measuring:

$$SAD_{dec} = \sum_{(x,y)\in B} |X_{2i-1}(x+dx, y+dy) + - X_{2i+1}(x-dx, y-dy)|, \quad (3)$$

   where the motion vector with components (dx, dy) is applied to block B.

2. Spatial smoothness can be estimated also at the decoder, in a similar way as at the encoder, this time computing the variance of the block in the motion-interpolated side information.

$$\sigma^2_{dec} = \frac{1}{M}\left[\sum_{(x,y)\in B} |Y_{2i}(x,y)|^2 - \left(\sum_{(x,y)\in B} Y_{2i}(x,y)\right)^2\right] \quad (4)$$

As before, if $SAD_{dec} \geq T^{TC}_{dec}$ or if $\sigma^2_{dec} \leq T^{SS}_{dec}$, or then the block is intra coded. Since the mode decision is computed at the decoder, the feedback channel already present in the coding architecture to perform rate control is also used to communicate the mode decision map to the encoder. With this solution, encoding and decoding are intertwined, as the encoder cannot start processing the frame X_{2i} until the decoder has computed the side information and sent the mode decisions to the encoder via the feedback channel. For this reason, this scheme can be adopted only in real time applications when encoding and decoding take place synchronously. The encoding/decoding process is the following:

1. Encoder: encode key frames X_{2i-1} and X_{2i+1}.
2. Decoder: decode key frames X_{2i-1} and X_{2i+1}.
3. Decoder: build the side information Y_{2i}.
4. Decoder: perform mode decision based on X_{2i-1} and X_{2i+1} according to (3) and (4).
5. Decoder: encode and send the mode decision map to the encoder through the feedback channel.
6. Encoder: decode the mode decision map and perform Wyner-Ziv/intra encoding accordingly.
7. Decoder: receive and decode the requested intra coded blocks.
8. Decoder: for the WZ-blocks, and for each bitplane, starting from the most significant one, correct the side information Y_{2i} to reconstruct X_{2i} by requesting parity bits and running the turbo decoding until the error probability is below a predefined threshold.

Comparing to the mode decision at the encoder, a better estimate of the temporal coherence is achieved but, on the other side, spatial smoothness cannot be computed on the original block to be encoded, thus giving a less significant metric.

4. EXPERIMENTAL RESULTS

Extensive experimental results have been performed using several test sequences in order to evaluate the effect of using block-based intra mode decision. In order to assess the advantages of the intra mode decision scheme, sequences characterized by different amount of motion have been included in the simulations, ranging from low motion (*News*, *Hall Monitor*), medium motion (*Coastguard*) and high motion (*Foreman*) sequences. QCIF resolution at 30 Hz has been used for these experiments. For blocks classified as Wyner-Ziv, the same tools and configuration parameters discussed in Section 2 are used. In particular, the motion-compensated side information is computed using lossy key frames encoded with H.263+ intra for a QP equal to 13, 10, 8, 5, respectively, depending on the number of decoded Wyner-Ziv bitplanes. Figure 2 shows the rate-distortion curves with the average PSNR computed across the whole sequence. Four coding schemes are compared:

- IST-PDWZ codec; side information computed from lossy key frames (1 out of 2 frames); all blocks in WZ frames encoded in Wyner-Ziv mode
- IST-PDWZ codec; side information computed from lossy key frames; part of the blocks in WZ frames are encoded in intra mode according to the mode decision carried out at the decoder
- IST-PDWZ codec; side information computed from lossy key frames; part of the blocks in WZ frames are encoded in intra mode according to the mode decision carried out at the encoder

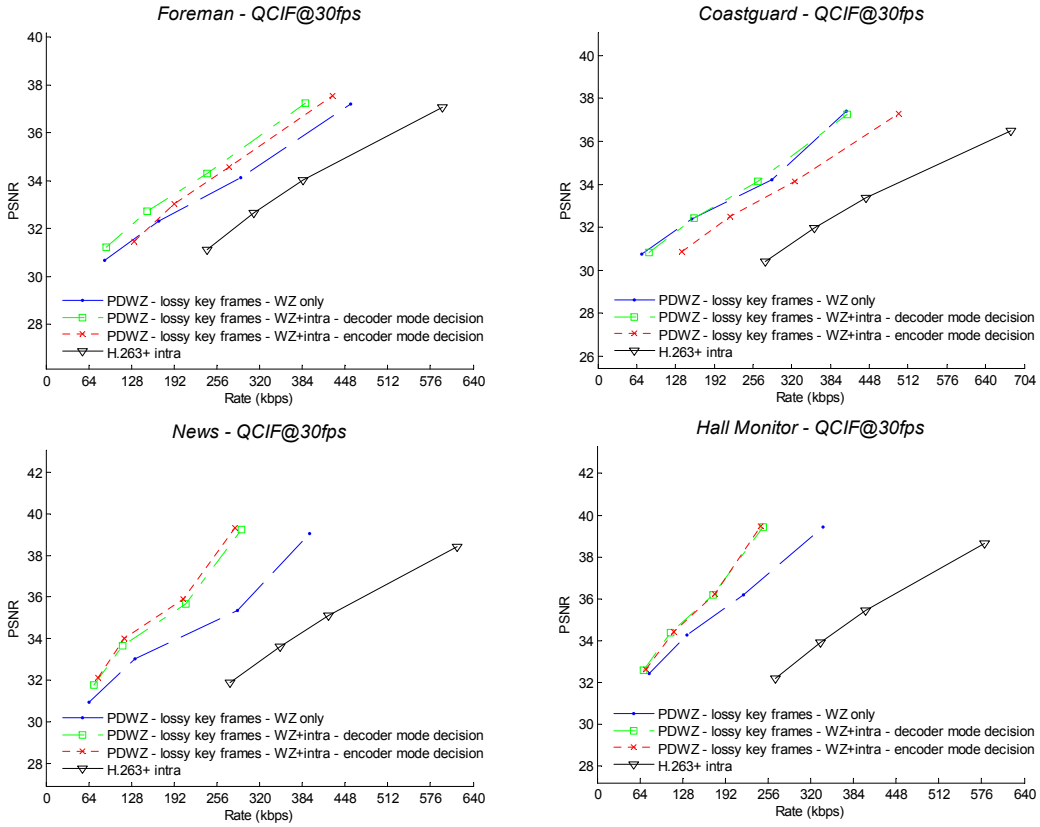

Fig. 2. IST-PDWZ rate-distortion curves: intra-mode decision at the encoder vs. at the decoder

- H.263+ intra coding

The quantization step size of the intra coded blocks in the Wyner-Ziv frames is set to be the same as for the key frames used to generate the corresponding side information. In sequences characterized by low motion, both intra mode decision schemes achieve nearly the same coding efficiency, with a slight gain for the encoder side mode decision for the *News* sequence. The latter fact can be justified observing that the temporal coherence is correctly estimated in both cases, whereas the spatial smoothness estimate is more reliable when computed at the encoder, as it has access to the original frame. On the other hand, when the amount of motion becomes more significant, the mode decision scheme at the decoder outperforms the one at the encoder. In fact, in this case, simple camera panning (as in *Coastguard*) or more complex motion (as in *Foreman*) lead to underestimate the quality of the side information when temporal coherence is computed at the encoder. In other words, a simple zero-motion frame difference as expressed by the SAD_{enc} metric is not a reliable indicator of the correlation with the actual side information. With respect to the case when all the blocks are Wyner-Ziv encoded, introducing an intra mode decision scheme gives a large coding efficiency gain, as much as 5 dB on average for the News sequence at high bitrates. Part of the observed gain is due to the fact that regions where the motion interpolation fails are now intra coded, as temporal redundancy cannot be exploited by the Wyner-Ziv codec. Also, since intra coding is performed in the transformed DCT domain, intra-frame spatial redundancy is successfully exploited.

5. CONCLUSIONS

In this paper we describe an intra mode decision scheme for frame-based pixel domain Wyner-Ziv video coding. We show that it is possible to achieve significant coding efficiency gains when the motion-compensated interpolation fails. Our future work will integrate the intra mode decision into a DCT-based Wyner-Ziv codec.

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] 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.

[3] A. Trapanese, M. Tagliasacchi, S. Tubaro, J. ao 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. ao 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] 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.