DYNAMIC KEY BLOCK DECISION WITH SPATIO-TEMPORAL ANALYSIS FOR WYNER-ZIV VIDEO CODING

Dung-Chan Tsai, Chang-Ming Lee, Wen-Nung Lie

Department of Electrical Engineering
National Chung Cheng University, Chia-Yi, Taiwan

ABSTRACT

Wyner-Ziv coding has been recognized as the most popular method up to now. For traditional WZC, side information is generated from intra-coded frames for use in the decoding of WZ frames. The unit for intra-coding is a frame and the distance between key-frames is kept constant. In this paper, the unit for intra-coding is a block, and the temporal distance between two consecutive key blocks can varying with time. A block is assigned a mode (WZ or intra-coded), depending on the result of spatio-temporal analysis, and encoded in an alternative manner. This strategy improves the overall coding efficiency, while maintaining a low encoder complexity. The performance gain can achieve up to 6 dB with respect to the traditional pixel-domain WZC.

Index Terms—Wyner-Ziv coding, distributed video coding, LDPC.

1. INTRODUCTION

Today's video coding standards, such as MPEG-X, H.26X, etc., are based on predictive coding techniques which use motion estimation (ME) to eliminate temporal redundancy. The complexity of this kind of video codec is high due to ME process. These techniques are suitable for applications where information is encoded only once and decoded many times. However, some applications require low complexity at the encoder side, but possibly disregard high complexity at the decoder side, such as wireless sensor networks. Two theoretical results show that it is possible to fit the above requirements: Slepian-Wolf theorem [1] and Wyner-Ziv (WZ) theorem [2]. The Slepian-Wolf theorem illustrates that given two correlated sources X and Y, the rate of independent and lossless encoding of X and Y is greater than or equal to the rate of joint encoding of X and Y, which means the coding efficiency of the joint encoding technique is higher. On the other hand, the Wyner-Ziv theorem is a lossy version of the Slepian-Wolf theorem, exhibiting the rate region of encoding X, given the distortion bound and without the knowledge of Y, but decoding X with side information related to Y.

When the above two algorithms are adopted in video coding, it is usually called distributed video coding (DVC). The pixel-domain DVC has been proposed in [3]. The conventional DVC architecture is based on a predefined group of picture (GOP). Frames in a video are organized as I-WZ-I-WZ…, where the GOP size equals 2. I frames are also recognized as the key frames to be intra-coded (as the source Y), which are responsible for generating the side information for WZ frames (as the source X) at the decoder side. With more accurate side information, the bit required to decode X could be reduced and the coding efficiency will be better. Therefore, some researchers focused on improving the side information estimation for X [4-6].

However, the temporal correlation between consecutive frames is not stationary and thus the fixed GOP structure will not be efficient enough. Therefore, a method to overcome the shortcoming of the traditional scheme is proposed in this paper. In our proposed scheme, each block in a frame is encountered a spatio-temporal analysis to be identified as a key or a WZ block (note that here we identify key blocks, instead of key frames, for the purpose of intra-coding). It results in a dynamic temporal distance (called temporal group of blocks (TGOBs)) between two consecutive key blocks. For blocks in the video sequence of still backgrounds, the TGOB size is large, while for blocks of moving objects, the TGOB size is small. We propose a new coding scheme based on this concept, while still keeping low encoding complexity.

2. KEY-BLOCK-BASED WYNER-ZIV CODING

In [7], the I-WZ-I-WZ… coding structure is maintained, but blocks of the WZ frames may be incurred a change into intra-coding, meaning that the WZ frames may be encoded in mixed modes. On the other hand, blocks of the key frames are all intra-coded. The gain of this coding-mode change for WZ frames is not high enough. On the other hand, Ascenso *et. al.* [8] made the GOP size (i.e., the distance between two consecutive key frames) dynamically adjustable, according to the motion activity along the sequence. However, these two techniques did not consider the tiny content variations in both the spatial and temporal directions.

These issues motivate us to explore the spatial and temporal correlation in a finer granularity. In the spatial direction, a block is adopted as the unit for intra- or WZ-encoding decision. In the temporal direction, the GOP

structure/length is dynamic, depending on the variations of temporal correlation in the video contents.

Figure 1 (placed in the last page) illustrates the proposed scheme based on the conventional pixel-domain DVC architecture. At the DVC encoder, the coding mode for each block is determined by the *block-mode decision unit,* which is then recorded in the *block-mode map queue*. If an intra mode is decided, the non-predictable intra coding of H.264 is realized. Otherwise, the WZ coding (WZC) is fulfilled. To be compatible to the conventional WZC process, a WZ frame generator manages the identified WZ blocks to form a full frame, before coding, by replacing the key blocks with zeros. As tradition [4], the WZC consists of a uniform quantizer, a lower-density parity-check (LDPC) encoder, buffer, a feedback channel and an LDPC decoder.

At DVC decoder, the intra bit-stream is decoded to reconstruct the key blocks, which are then sent to the decoded key-block queue. As tradition, the decoded key blocks are used by the side information generator to interpolate the WZ blocks along the temporal direction. The interpolated/estimated WZ blocks are recognized as the side information $\hat{X}$ for LDPC-decoding X. By integrating the reconstructed key and WZ blocks from both decoders, the reconstructed frame can be completed.

3. DETAILS OF THE PROPOSED SCHEME

Traditionally, side information, generated by interpolating between two consecutive key frames (separated apart at a GOP length), is used for LDPC-decoding a WZ frame. If the side information is accurate enough, meaning that $\hat{X}$ and X are approximate to each other, a less number of parity bits for LDPC decoding is required and a better RD performance can be achieved.

However, as mentioned above, our policy for intra-coding is fine granularity in both the spatial and temporal directions, resulting in a compromise between the accurate estimation of the WZ information and the transmission bit-rate. Let the structure of "IB-WZB…WZB-IB-WZB-…." (IB: intra-coded block, WZB: WZ-coded block) in the temporal direction be named Temporal Group of Blocks (TGOBs). The TGOB size is large for still background areas and small for dynamic foreground regions. Notice that TGOB size is dynamic and time-varying, according to the video contents. For example, in Fig.2, for the $Block_{0,1}$, the TGOB size varies in the order of 3, 1, 2 and 1.

3.1. Block-mode decision unit

In our system, each frame is divided into non-overlapping blocks, each is of 16×16 pixels. Each block is categorized into a key-block which is intra-coded or a WZ-block which is quantized in the pixel-domain and LDPC-coded. All blocks in the first frame are intra-coded and serve as the initial key blocks along the sequence.

The block-mode is determined based on the temporal correlation with the previous key block at the same position and on the spatio correlation within a block. Two criteria in [7] are adopted:

$$SAD = \sum_{(x,y)\in \boldsymbol{Block}_{i,j}} \left| I^{i,j,t}(x,y) - I^{i,j,t-d_1}(x,y) \right| \quad (1)$$

$$\sigma^2 = \frac{1}{S}\sum_{(x,y)\in Block_{i,j}} \left| I^{i,j,t}(x,y) \right|^2 - \left(\frac{1}{S}\sum_{(x,y)\in Block_{i,j}} I^{i,j,t}(x,y) \right)^2 \quad (2)$$

where t is the time index, (i, j) indicates the block index, (x,y) is the pixel coordinate, d_1 is the distance to the previous key block (in the temporal direction), S is the size of a block, and all *I's* represent the input video data. If *SAD* is above a pre-determined threshold or σ is below the other, that block is identified as a key-block and then intra-coded. Otherwise, it is WZ-coded. However, a limiting upper bound U of TGOB size would be used to avoid a long decoding delay and accumulated inaccuracy. The block-mode map for each frame is created and transmitted to the decoder.

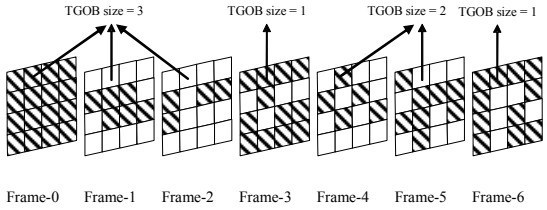

Fig. 2. The concept of dynamic TGOB, where the shaded blocks represent the key blocks and the others are WZ blocks.

3.2. WZ frame generator

Since the number of WZ blocks for each frame is not kept a constant, different sizes of LDPC parity check matrix are required. To solve this annoying problem, the WZ frame generator provides a full frame by replacing the identified key blocks with zeros. The inserted zeros will definitely make the size of the LDPC parity check matrix fixed, sacrifice coding efficiency, but reduce the decoding complexity (zeros will significantly constrain the possible paths during decoding). Accordingly, the block-mode map should be required when performing LDPC decoding.

3.3. LDPC coding and decoding

After WZ frame generation, the resulting frame is fed into a uniform scalar quantizer with 2^M levels, where M bit-planes are extracted. Each bit-plane is sent to an LDPC encoder, where the produced bit-streams are temporarily stored in an output buffer and sent on-demand to the decoder. At decoder side, the side information generator (discussed later) exploits the decoded key-blocks to obtain the estimates (i.e., the side information) of the WZ blocks. The side information is then used by the iterative LDPC decoder for initial probability estimates for each bit-plane

reconstruction. The WZ bit-streams of the higher significant bit-planes are decoded first. After decoding a part of or all bit-planes, pixels of the WZ blocks can be recovered to some fidelity.

3.4. Side information generator

The method of linear interpolation is used to estimate the WZ blocks from the decoded key-blocks. Let d_1 and d_2 be the temporal distances of the current block to the preceding and successive key-blocks, respectively. Then pixels of a WZ block $WZB^{i,j,t}$ are estimated by:

$$\hat{WZB}^{i,j,t}(x,y) = \frac{d_1 \cdot \hat{IB}^{i,j,t-d_1}(x,y) + d_2 \cdot \hat{IB}^{i,j,t+d_2}(x,y)}{d_1 + d_2}, \quad (3)$$

where $\hat{IB}^{i,j,t}(x,y)$ represents the decoded key-block data. A collection of $\{\hat{WZB}^{i,j,t}(x,y)\}$ then forms the side information in decoding the corresponding WZ bit-stream. Note that the side information also includes the block-mode map, which helps the system to insert zero blocks at key-block positions (recall that the key-blocks are replaced with zeros to form a WZ frame).

3.5. LDPC buffer and decoded key block queue

To decode a WZ block, the side information should be available first. The preceding and successive key blocks, separated $d_1 + d_2$ (i.e., TGOB size) frames apart, have to be completely decoded and available at the decoder buffer. Clearly, this is a non-causal system and a system delay is unavoidable. To avoid varying time delays for WZ blocks, an extended LDPC buffer of U frames is adopted. Accordingly, a decoded key-block queue of ($2U$-1) frames long is designed at the decoder to guarantee the availability of the preceding and successive key-blocks for interpolation.

Fig. 3 illustrates the relative timing for components in our proposed WZ-Coding system, where the frame number under processing/storage is indicated on the top of the box. For example, letting the processed frame at intra decoder be Frame-t, the decoded key-block queue would contain the decoded key-blocks for frames (t-$2U$+1) to (t-1), the LDPC decoder would process frame-(t-U+1), and the encoder buffer would contain bit-streams for frames (t-U+1) to t. At this moment, the integration of the decoded key and WZ blocks would form the output for frame-(t-U+1). In this way, the largest delay for decoding a WZ block is U-1 frames.

Note that the TGOB lengths for WZ blocks are different, that is, to decode the WZ blocks for a given frame, the sets of two key-blocks for interpolation may not be synchronously available. However, with a decoded key-block queue of ($2U$-1) frames long, this availability/ synchronization can be guaranteed. The size of this queue depends on how many key-blocks are identified/allowed during a ($2U$-1)-frame of interval. In contrast to the decoded key-block queue, the LDPC encoder buffer requires a length of U frames only. Sure, the buffer size depends on the length of the LDPC bit-streams accumulated during the past U frames of interval.

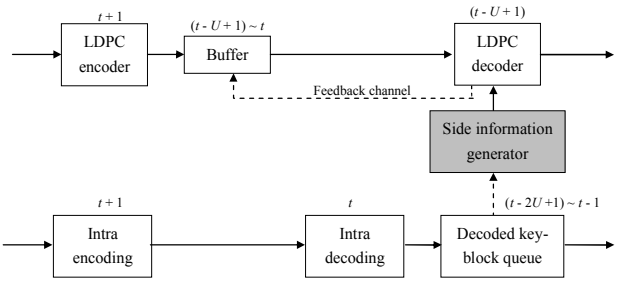

Fig. 3. Relative timing for components in our proposed WZ coding system.

4. EXPERIMENT RESULTS

In our proposed WZ coding scheme, the parameters include: number (M) of bit-planes requested by the LDPC decoder, lower (L) and upper (U) bounds of the TGOB size, thresholds for SAD in (1) and σ in (2), and the quantization parameter (QP) for intra-coding. Importantly, we set M=1, L=1, and U=10. The test sequences are all of QCIF format. The identified key blocks are encoded with the non-predictable intra-coding mode in H.264/AVC standard.

Our coding scheme is compared to the LDPC-based traditional one [3] which adopts the "I-WZ-I-WZ…." coding structure (i.e., GOP size = 2) and a similar weighted-average interpolation, as expressed in (3), for side information generation. The bit-rates and PSNRs are averaged over the whole sequence. The simulation results for the first 200 frames of the "Salesman" and the "Akiyo" sequences are shown in Fig. 4. As it can be seen from the plot, the proposed WZC scheme has a PSNR gain by 2 to 6 dB at the same bit-rate. This gain is even up to 11 dB, compared to the H.264 intra coding in "I-I-I…" structure.

The reason of the gains is explained as follows. For the conventional WZ codec, the TGOB size is kept a constant 2. This is not efficient for videos of static contents. We need a codec which is capable of adapting to dynamic video contents. The TGOB size should be smaller for lower temporal correlation, and larger for higher temporal correlation. Fig.5 shows the histogram of the TGOB size for the "Salesman" sequence. Obviously, a large portion of blocks are identified with larger TGOBs (nearly 41% for TGOB size > 5 frames).

5. CONCLUSIONS AND FUTURE WORK

In this paper, a new key-block-based WZ coding scheme is proposed. Dynamic key-block decision adapts our system to varying video contents and getting more coding efficiency. However, if the simple weighted-interpolation is replaced with the MCI technique for side information generation, the

performance is expected to be further improved. Our future work will involve realizing the rate control issue and improving the channel coding performance.

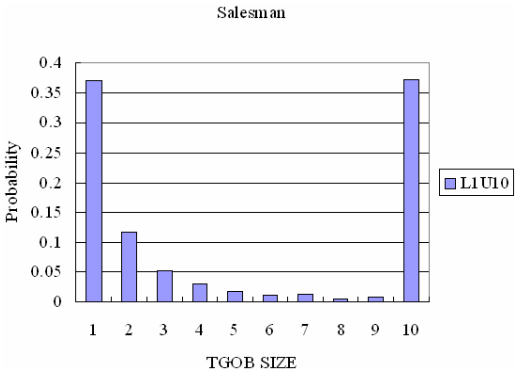

Fig. 5. Histogram of TGOB sizes ($L = 1$, $U = 10$) for the "Salesman" sequence.

REFERENCE

[1] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. on Information Theory*, Vol. 19, no. 4, p.p. 471-480, July 1973.

[2] D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. on Information Theory*, Vol. 22, pp. 1-10, Jan. 1976.

[3] A. Aaron, R. Zhang and B. Girod, "Wyner-Ziv Coding of Motion Video," *36th Asilomar Conference on Signals, Systems and Computer*, Pacific Grove, USA, November 2002.

[4] J. Ascenso, C. Brites, and F. Pereira, "Motion Compensated Refinement for Low Complexity Pixel Based Distributed Video Coding," *IEEE Int'l Conf. on Advanced Video and Signal-Based Surveillance*, Como, Italy, Sep. 2005.

[5] D. Kubasov, and C. Guillemot, "Mesh-based motion-compensated interpolation for side information extraction in distributed video coding," *IEEE Int'l Conf. on Image Processing* (ICIP), Atlanta, USA, October 8-11, 2006.

[6] J. Ascenso, C. Brites, and F. Pereira, "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, June 29 – July 2, 2005.

[7] M. Tagliasacchi, A. Trapanese, S. Tubaro, J. Ascenso, C. Brites, and F. Pereira, "Intra Mode Decision Based On Spatio-Temporal Cues In Pixel Domain Wyner-Ziv Video Coding", *IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing* (ICASSP), Toulouse, France, May 14-19, 2006.

[8] J. Ascenso, C. Brites, F. Pereira, "Content Adaptive Wyner-Ziv Video Coding Driven by Motion Activity," *IEEE Int'l Conf. on Image Processing* (ICIP), USA, October 8-11, 2006.

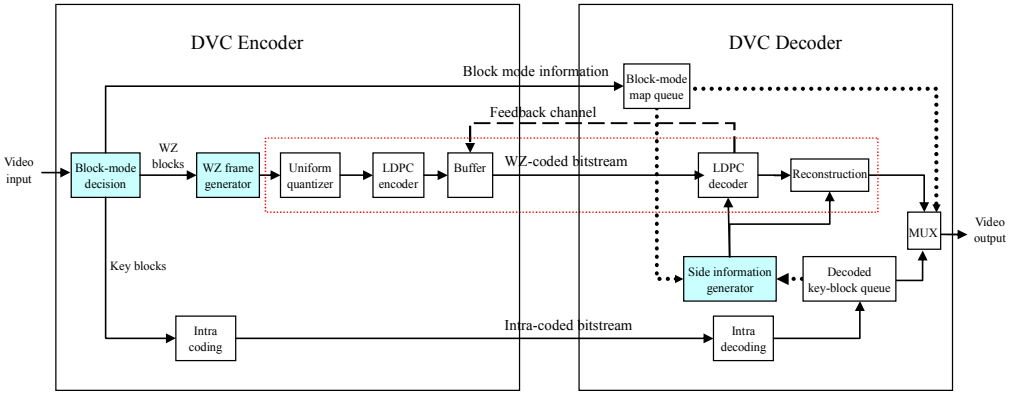

Fig. 1. Proposed key-block-based Wyner-Ziv coding architecture.

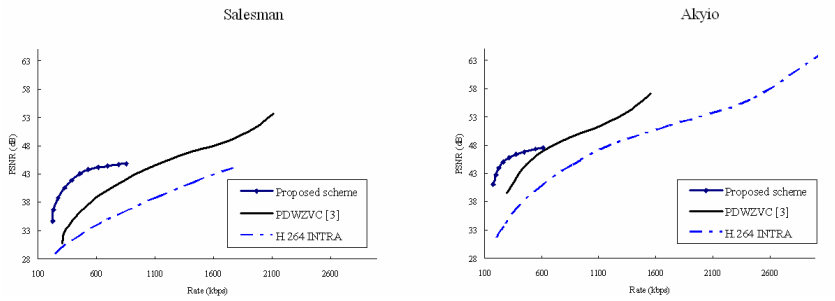

Fig. 4. RD performances for the "Salesman" and "Akiyo" sequences (each is of 200-frame long).