WYNER-ZIV VIDEO CODING: A REVIEW OF THE EARLY ARCHITECTURES AND FURTHER DEVELOPMENTS

Fernando Pereira[1], Catarina Brites[1], João Ascenso[2], Marco Tagliasacchi[3]

[1] Instituto Superior Técnico – Instituto de Telecomunicações, Lisbon, Portugal
E-mail: {fernando.pereira, catarina.brites}@lx.it.pt
[2] Instituto Superior de Engenharia de Lisboa – Instituto de Telecomunicações, Lisbon, Portugal
E-mail: joao.ascenso@lx.it.pt
[3] Politecnico di Milano – Dipartimento di Elettronica e Informazione, Milan, Italy
E-mail: marco.tagliasacchi@polimi.it

ABSTRACT *

In 2002, the video coding community faced the emergence of a new video coding paradigm, the so-called Wyner-Ziv video coding, which was represented by two early solutions designed by the Stanford University and the University of California, Berkeley research teams. This paper intends to briefly review, and compare these two early Wyner-Ziv video coding solutions, notably from the functional point of view. Moreover, this paper reviews some important developments of the Stanford Wyner-Ziv coding architecture, which has become the most popular in the literature.

Index Terms — distributed video coding, Wyner-Ziv video coding, coding efficiency, low complexity, error resilience, scalability

1. INTRODUCTION

The main objective of digital video coding technologies is to compress the original data into a much smaller number of bits, while preserving an acceptable video quality. These technologies are behind the success and rapid deployment of products and services such as digital cameras, digital television, and DVDs, among others. Most available video coding standards, notably the ITU-T H.26X and ISO/IEC MPEG-X families of standards, adopt the so-called predictive video coding paradigm where the temporal and spatial correlations are exploited at the encoder by using a motion compensated prediction loop and a spatial transform, respectively. As a consequence, this video coding solution typically leads to rather complex encoders and much simpler decoders, with a rigid allocation of the complexity between the transmitter and the receiver. This approach fits well some application scenarios, e.g. broadcasting, where a few (complex) encoders provide coded content for millions of (simpler) decoders.

With the wide deployment of mobile and wireless networks, there is a growing number of applications where many senders deliver data to a central receiver. Typically, these emerging applications require light encoding complexity, high compression efficiency, robustness to packet losses and, often, also low latency/delay. To address some of these issues, some research groups revisited the video coding problem at the light of an Information Theory result from the 70s: the Slepian-Wolf theorem [1]. According to this theorem, the minimum rate needed to independently encode two statistically dependent discrete random sequences, X and Y, is the same as for joint encoding. While the Slepian-Wolf theorem deals with lossless coding, in 1976, A. Wyner and J. Ziv studied the case of lossy coding with side information (SI) at the decoder. Under some hypothesis on the joint statistics, the Wyner-Ziv theorem [2] states that when the side information (i.e. the correlated source Y) is made available only at the decoder there is no coding efficiency loss in encoding X, with respect to the case when joint encoding of X and Y is performed. The Slepian-Wolf and the Wyner-Ziv theorems suggest that it is possible to encode two statistically dependent signals independently and decoding them jointly, while approaching the coding efficiency of conventional predictive coding schemes, which rely on joint encoding and decoding instead. The new coding paradigm, known as Distributed Video Coding (DVC) avoids the computationally intensive temporal prediction loop at the encoder, by shifting the exploitation of the temporal redundancy at the decoder. This is a significant advantage in a large range of emerging application scenarios, including wireless video cameras, wireless low-power surveillance, video conferencing with mobile devices, and visual sensor networks.

With the theoretical doors opened, the practical design of Wyner-Ziv (WZ) video codecs, a particular case of DVC, started around 2002, following important developments in channel coding technology. The first practical WZ solutions have been developed at Stanford University [3,4] and UC Berkeley [5,6]. As of today, the most popular WZ video codec design in the literature is clearly the Stanford architecture, which works at the frame level and is characterized by a feedback channel based decoder rate control. On the other hand, the Berkeley architecture, known as PRISM (Power-efficient, Robust, hIgh compression Syndrome based Multimedia coding), works at the block level and is characterized by an encoder side rate controller based on the availability of a reference frame.

Due to their popularity in the research community and the major technical evolution in recent years, Section 2 presents and compares from the functional point of view these two WZ video coding architectures. In Section 3, a brief review of some of the architectural developments which derived from the initial Stanford architecture is given. Due to space constraints, this paper will only address monoview video coding. Multiview video coding has also been addressed in the literature; for a complete review, see [7].

2. THE EARLY WYNER- ZIV VIDEO CODING ARCHITECTURES

This section introduces and compares the Stanford and Berkeley Wyner-Ziv video coding architectures.

2.1 The Stanford WZ Video Coding Architecture

The Stanford WZ video coding architecture was first proposed in 2002 for the pixel domain [3] and later extended to the transform domain [8] where DCT coefficients are WZ coded. In summary, the (more efficient) transform domain WZ video codec shown in Fig. 1 works as follows:

Splitting Frames: The video sequence is divided into Wyner-Ziv (WZ) frames and key frames. The key frames are encoded in intra-frame mode, e.g. using H.263+ Intra or H.264/AVC Intra, and inserted periodically determining the GOP size.

* The work presented was developed within VISNET II, a European Network of Excellence (http://www.visnet-noe.org).

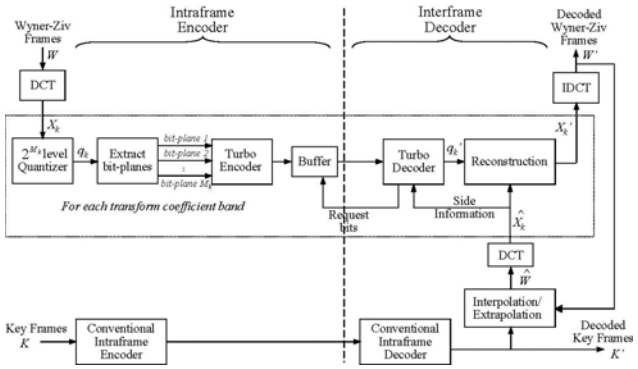

Fig. 1 – Stanford WZ video coding architecture [8]

Transform: A block-based transform, typically a DCT, is applied to each WZ frame. The DCT coefficients of the entire WZ frame are then grouped together, according to the position occupied by each DCT coefficient within a block, forming DCT coefficient bands.

Quantization: Each DCT band is uniformly quantized with a number of levels that depends on the target quality [8]. For a given band, bits of the quantized symbols are grouped together, forming bitplanes, which are then independently turbo encoded.

Turbo Encoding: The turbo encoding of each DCT band starts with the most significant bitplane (MSB). The parity information generated for each bitplane is then stored in the buffer and sent in chunks/packets upon decoder requests, made through the feedback channel.

Side Information Creation: The decoder creates the side information for each WZ coded frame, by performing a motion-compensated frame interpolation (or extrapolation) using the closest already decoded frames. The side information for each WZ frame is taken as an estimate (noisy version) of the original WZ frame. The better it is the estimate, the smaller is the number of 'errors' the turbo decoder has to correct and the bitrate needed.

Correlation Noise Modeling: The residual statistics between corresponding coefficients in the WZ frame and the side information is assumed to be modeled by a Laplacian distribution whose parameter was initially estimated using an offline training phase.

Turbo Decoding: Once the side information DCT coefficients and the residual statistics for a given DCT coefficients band are known, each bitplane is turbo decoded (starting from the MSB one). The turbo decoder receives from the encoder successive chunks of parity bits following the requests made through the feedback channel. To decide whether or not more bits are needed for successful decoding of a certain bitplane, the decoder uses a request stopping criterion. After successfully turbo decoding the MSB bitplane of a DCT band, the turbo decoder proceeds in an analogous way with the remaining bitplanes associated to the same band. Once all the bitplanes of a DCT band are successfully turbo decoded, the turbo decoder starts decoding the next band.

Reconstruction: After turbo decoding all the bitplanes associated to each DCT band, the bitplanes are grouped together to form the decoded quantized symbol stream associated to each band. Once all decoded quantized symbols are obtained, it is possible to reconstruct the matrix of DCT coefficients. The DCT coefficients bands for which no WZ bits were transmitted are replaced by the corresponding DCT bands of the side information.

Inverse Transform: After all DCT bands are reconstructed, a block-based inverse transform, typically the IDCT, is performed and the decoded WZ frame is obtained.

Frame Reordering: Finally, to get the decoded video sequence, decoded key frames and WZ frames are conveniently mixed.

Over the last few years, many improvements have been proposed for most of the modules in the initial Stanford WZ video codec: e.g. LDPC codes instead of turbo codes [9,10], better side information estimation [11], dynamic correlation noise modeling [12], enhanced reconstruction [13], etc. Other proposed solutions required revisiting the original architecture by introducing major changes, e.g. selective Intra coding of blocks in the WZ frame [14], selective transmission of hash signatures by the encoder [15,16], removal of the feedback channel [17], provision of scalability [18,19,20] and error resilience features [21,22,23], etc.

2.2 The Berkeley WZ Video Coding Architecture

Almost at the same time of the Stanford WZ coding solution, another WZ video coding approach has been proposed at UC Berkeley, known in the literature as PRISM [5,6]. In summary, the PRISM codec is shown in Fig. 2 and it works as follows:

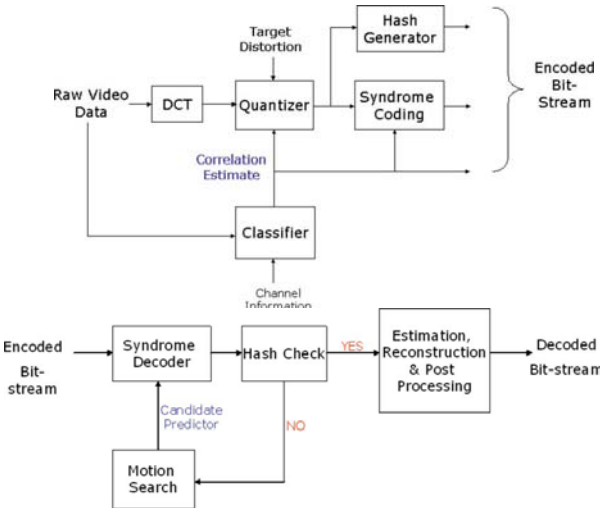

Fig 2 – PRISM encoder and decoder architectures [6]

Transform: Each video frame is divided into 8×8 samples blocks and a DCT is applied over each block.

Quantization: A scalar quantizer is applied to the DCT coefficients corresponding to a certain target quality.

Classification: Before encoding, each block is classified into one of several pre-defined classes depending on the correlation between the current block and the predictor block in the reference frame. Depending on the allowed complexity at the encoder, such a predictor can be either the co-located block, or a motion-compensated block [6]. The classification stage decides the coding mode for each block of the current frame: no coding (skip class), traditional Intraframe coding (entropy coding class) or syndrome coding (syndrome coding classes), depending on the estimated temporal correlation. The blocks classified in the syndrome coding classes are coded using a WZ coding approach as described below. The coding modes are then transmitted to the decoder as header information.

Syndrome Coding: For those blocks that fall in the syndrome coding classes, only the least significant bits of the quantized DCT coefficients in a block are encoded, since it is assumed that the most significant bits can be inferred from the side information. The number of least significant bits to be sent to the decoder depends on the syndrome class the block belongs to. Within the least significant bits, the lower part is encoded using a *(run, depth, path, last)* 4-tuple based entropy codec. The upper part of the least significant bits is coded using a coset channel code, in this case a BCH code, since it works well for small-block lengths as it is the case here.

Hash Generation: In addition, for each block, the encoder sends a 16-bit CRC checksum as a signature of the quantized DCT coefficients. This is needed in order to select the best candidate block (SI) at the decoder as explained below.

Motion Search: The decoder generates side information candidate blocks, which correspond to all half-pixel displaced blocks in the reference frame, in a window around the block to be decoded.

Syndrome Decoder: Each of the candidate blocks plays the role of side information for syndrome decoding, which consists in two steps: one step deals with the entropy coded least significant bitplanes and the other step with the coset channel coded bitplanes.

Hash Checking: Each candidate block leads to a decoded block, from which a hash signature is generated. In order to detect successful decoding, the latter is compared with the CRC hash received from the encoder. Candidate blocks are visited until decoding leads to hash matching.

Reconstruction and IDCT: Once the quantized sequence is recovered, it is used along with the corresponding side information to get the best reconstructed block. The minimum mean squared estimate is computed from the side information and the quantized block.

2.3 Comparing the Early WZ Video Coding Solutions

While the reasons for the research community to have adopted more enthusiastically the Stanford architecture are not fully clear, it was very likely a relevant factor that more literature was available, and overcoming the initial implementation barrier was easier. From the technical point of view, the following main functional differences may be highlighted (Stanford versus Berkeley):

1. *Frame based versus block based coding*. In the latter approach, it is easier to accommodate coding adaptability to address the highly non-stationary statistics of video signals.
2. *Decoder rate control versus encoder rate control*. In the former case, a feedback channel is needed, restricting the scope to real-time applications.
3. *Very simple encoder versus smarter encoder*. Enabling limited inter-frame operations at the encoder allows incorporating spatially varying coding mode decisions. For example, acknowledging that it is useless to adopt a WZ coding approach when the correlation is too weak or inexistent.
4. *More sophisticated channel codes*, notably turbo codes and later LPDC codes, *versus simpler channel codes*, e.g. BCH codes.
5. *No auxiliary data versus hash codes* sent by the encoder to help the decoder in the motion estimation process.
6. *Less intrinsically robust to error corruption versus higher resilience to error corruption* due to the PRISM motion search like approach performed at the decoder.

With time, some of the differences above have disappeared, e.g. there are nowadays Stanford based coding solutions with selective block based Intra coding [14], encoder transmitted hash signatures [15,16], and without feedback channel [17].

However, after a few years, the performance gap between the two early solutions seems to be rather significant. In November 2007, the European project DISCOVER published error free rate-distortion (RD) performance results for a WZ video codec based on the Stanford architecture which is able to outperform H.264/AVC Intra and sometimes even H.264/AVC 'zero-motion' standard coding [9,10]. In October 2007, the Berkeley team published error free RD performance results which only slightly outperform H.263+ coding [6].

3. DEVELOPMENTS ON THE STANFORD WYNER-ZIV VIDEO CODING ARCHITECTURE

In recent years, a significant number of research groups around the world have adopted the Stanford WZ coding architecture and changed it to address certain needs and functionalities. Due to space limitations, this section will just briefly describe some examples of possible architectural variations of the initial Stanford WZ video coding architecture; for the same reason, only a very limited number of references are included.

3.1 Improving Coding Efficiency

Because the initial RD performance was rather poor in comparison with the alternative solutions provided by the available standards, e.g. H.263+ and H.264/AVC, most of the research has focused on improving the coding efficiency in the context of low complexity encoding.

Selective Block based Intra Coding

Somehow inspired by the PRISM approach, the addition of a block based classification module to the WZ video encoder, allowing to select a coding mode adapted to the available temporal correlation has been proposed in [14]. A mode decision scheme (applied either at the encoder or at the decoder) works in such a way that when the estimated correlation is weak, intra coding is performed on a block-by-block basis. Both spatial and temporal criteria are used to determine whether a block is better intra coded or not. With respect to the case when all the blocks are WZ encoded, introducing an intra mode decision scheme gives as much as 5 dB, on average, for the News sequence at high bitrates.

Encoder Hash Signatures for Better Side Information

Still trying to overcome the 'blind' frame based approach adopted by the initial Stanford WZ coding solution, and recognizing that the temporal correlation to exploit is not uniform within the frames, some researchers have changed the architecture to incorporate the capability for the encoder to send some hash signatures in order to help the decoder generating a better side information [15,16]. In [15], the hash code for an image block simply consists of a small subset of coarsely quantized DCT coefficients of the block. Since the hash requires fewer bits than the original data, the encoder is allowed to keep the hash codewords for the previous frame in a small hash store. Strictly speaking, the encoder is no longer intraframe due to the hash store. In [15], significant gains over conventional DCT-based intraframe coding are reported, while having comparable encoding complexity.

Advanced Side Information, Noise Correlation Modeling and Reconstruction

It is worthwhile to mention the Stanford based WZ video codec developed by the European Project DISCOVER [9,10], since it provides the best known RD performance, notably due to the sophisticated side information creation, the dynamic online noise correlation modeling, and the optimal reconstruction. The side information module uses block matching based on a modified mean absolute difference (MAD) to regularize the motion vector field; after, a hierarchical coarse-to-fine bidirectional motion estimation is performed (with half-pixel precision); spatial motion smoothing based on a weighted vector median filter is applied afterwards to the obtained motion field to remove outliers before motion compensation is finally performed. The correlation noise modeling is performed at the decoder at various levels of granularity, e.g. band or coefficient levels, allowing a dynamic adaptation of the model to the varying temporal correlation. Finally, the decoded values are reconstructed using an optimal MSE based approach using closed-form expressions derived for a

Laplacian correlation model [13]. In [10], gains over H.264/AVC 'zero-motion' for Coast Guard and gains over H.264/AVC Intra for most tested sequences are already reported.

3.2 Removing the Feedback Channel

The feedback channel is very likely the most controversial architectural element in the Stanford WZ video coding solution since it implies not only the presence of the feedback channel itself but also that the application works in real-time; the application and the video codec must be also able to accommodate the delay associated to the feedback channel. On the other hand, the usage of the feedback channel simplifies the rate control problem since the decoder, knowing the available side information, can easily adjust the necessary bitrate. To allow the Stanford solution to be applicable to other applications not fulfilling the conditions above, a variation performing encoder rate control (without feedback channel) is proposed in [17]. This paper reports a loss up to 1.2 dB, especially for the highest qualities, between the encoder and decoder rate control solutions.

3.3 Improving Error Resilience

Distributed video coding principles have been extensively applied in the field of robust video transmission over unreliable channels. WZ video codecs are characterized by in-built error robustness, due to the lack of the prediction loop that characterizes conventional motion-compensated predictive codecs. Most of the WZ video coding schemes that focus on error resilience try to increase the robustness of standard encoded video by adding redundant information encoded according to WZ video coding principles. One of the first works with this focus [21] uses auxiliary WZ encoded data sent only for some frames, to stop drift propagation at the decoder. In [4] an MPEG-2 Video coded bitstream is protected by a cascade of WZ bitstreams achieving graceful degradation with increasing channel error rate without using a scalable representation. The Systematic Loss Error Protection (SLEP) framework has been later extended to the case of H.264/AVC [22]. Finally, the error resilience performance of the feedback channel based transform domain WZ video codec by the Project DISCOVER [9,10] has been investigated in [23].

3.4 Providing Scalability

In current scalable codecs there is typically a predictive approach from lower layers to upper layers, requiring the encoder to use as reference the previous layers decoded frames in order to create the successive enhancements (SNR, spatial resolution). However, the WZ prediction loop free approach between the scalable layers does not require anymore a deterministic knowledge of the previous layers (just a correlation model) which means the layers may be generated by various, different and unknown codecs. In [18], a FGS (fine granularity scalability) WZ codec is proposed, where refinement bitplanes are encoded with a hybrid approach, using either LPDC codes or conventional VLC source coding tools. A layered WZ coding architecture is proposed in [19], achieving both scalability and error resilience. Finally, several WZ based scalable architectures providing different types of scalability are presented in [20].

4. FINAL REMARKS

It is presently more and more accepted that the Distributed Source Coding (DSC) principles are leading to varied tools which may help to solve different problems, e.g. coding, authentication and secure biometrics. While it is difficult to state, at this stage, if any video coding product will ever use DSC principles, and for what purpose, it is most interesting to study and research towards this possibility. Further WZ video coding research should address issues such as side information creation, iterative decoding, correlation noise modeling, novel channel codes, rate control, and WZ selective coding.

5. REFERENCES

1. J. Slepian and J. Wolf, "Noiseless Coding of Correlated Information Sources", IEEE Trans. on Information Theory, vol. 19, nº 4, pp. 471 - 480, July 1973.
2. A. Wyner and J. Ziv, "The Rate-Distortion Function for Source Coding with Side Information at the Decoder", IEEE Trans. on Information Theory, vol. 22, nº 1, pp. 1 - 10, January 1976.
3. A. Aaron, R. Zhang and B. Girod, "Wyner-Ziv Coding of Motion Video", Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, November 2002.
4. B. Girod, A. Aaron, S. Rane and D. Rebollo Monedero, "Distributed Video Coding", Proceedings of the IEEE, vol. 93, nº 1, pp. 71 - 83, January 2005.
5. R. Puri and K. Ramchandran, "PRISM: A New Robust Video Coding Architecture Based on Distributed Compression Principles", 40th Allerton Conference on Communication, Control and Computing, Allerton, USA, October 2002.
6. R. Puri, A. Majumdar and K. Ramchandran, "PRISM: A Video Coding Paradigm with Motion Estimation at the Decoder", IEEE Trans. on Image Process., vol. 16, nº 10, pp. 2436 - 2448, Oct. 2007.
7. C. Guillemot, F. Pereira, L. Torres, T. Ebrahimi, R. Leonardi and J. Ostermann, "Distributed Monoview and Multiview Video Coding", IEEE Signal Processing Magazine, vol. 24, nº 5, pp. 67 - 76, September 2007.
8. A. Aaron, S. Rane, E. Setton and B. Girod, "Transform-domain Wyner-Ziv Codec for Video", VCIP'04, San Jose, California, USA, January 2004.
9. X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov and M. Ouaret, "The DISCOVER Codec: Architecture, Techniques and Evaluation", PCS'07, Lisbon, Portugal, November 2007.
10. DISCOVER Page, http://www.img.lx.it.pt/~discover/home.html
11. J. Ascenso, C. Brites and F. Pereira, "Improving Frame Interpolation with Spatial Motion Smoothing for Pixel Domain Distributed Video Coding", EURASIP Conf. on Speech and Image Processing, Multimedia Communications and Services, Smolenice, Slovak Republic, June 2005.
12. C. Brites, J. Ascenso and F. Pereira, "Studying Temporal Correlation Noise Modeling for Pixel Based Wyner-Ziv Video Coding", ICIP'06, Atlanta, USA, October 2006.
13. D. Kubasov, J. Nayak and C. Guillemot, "Optimal Reconstruction in Wyner-Ziv Video Coding with Multiple Side Information", MMSP'07, Chania, Crete, Greece, October 2007.
14. A. Trapanese, M. Tagliasacchi, S. Tubaro, J. Ascenso, C. Brites and F. Pereira, "Intra Mode Decision Based on Spatio-Temporal Cues in Pixel Domain Wyner-Ziv Video Coding", ICASSP'06, Toulouse, France, May 2006.
15. Aaron, S. Rane and B. Girod, "Wyner-Ziv Video Coding with Hash-Based Motion Compensation at the Receiver", ICIP'04, Singapore, October 2004.
16. J. Ascenso and F. Pereira, "Adaptive Hash-Based Side Information Exploitation for Efficient Wyner-Ziv Video Coding", ICIP'07, San Antonio, TX, USA, September 2007.
17. C. Brites and F. Pereira, "Encoder Rate Control for Transform Domain Wyner-Ziv Video Coding", ICIP'07, San Antonio, Texas, USA, September 2007.
18. H. Wang, N. Cheung and A. Ortega, "A Framework for Adaptive Scalable Video Coding Using Wyner-Ziv Techniques", EURASIP Journal on App. Signal Processing Volume 2006, Article ID 60971
19. Q. Xu and Z. Xiong, "Layered Wyner-Ziv Video Coding", IEEE Trans. Image Processing, vol. 15, nº 12, pp. 3791-3803, Dec. 2006.
20. M. Ouaret, F. Dufaux and T. Ebrahimi, "Codec-Independent Scalable Distributed Video Coding", ICIP'07, San Antonio, TX, USA, September 2007.
21. A. Sehgal, A. Jagmohan, and N. Ahuja, "Wyner-Ziv Coding of Video: an Error-Resilient Compression Framework", IEEE Transactions on Multimedia, vol. 6, nº 2, pp. 249–258, April 2004.
22. S. Rane, P. Baccichet, and B. Girod, "Modeling and Optimization of a Systematic Lossy Error Protection System based on H.264/AVC Redundant Slices", PCS'06, Beijing, China, April 2006

23. J. Pedro et al., “Studying Error Resilience Performance for a Feedback Channel based Transform Domain Wyner-Ziv Video Codec”, PCS’07, Lisbon, Portugal, Nov. 2007.