DISTRIBUTED VIDEO CODING WITH MULTIPLE SIDE INFORMATION

Xin Huang[1]*, Catarina Brites*[2]*, João Ascenso*[3]*, Fernando Pereira*[2]*, and Søren Forchhammer*[1]

[1]DTU Fotonik, Technical University of Denmark
[2]Instituto Superior Técnico - Instituto de Telecomunicações, Portugal
[3]Instituto Superior de Engenharia de Lisboa - Instituto de Telecomunicações, Portugal

ABSTRACT

Distributed Video Coding (DVC) is a new video coding paradigm which mainly exploits the source statistics at the decoder based on the availability of some decoder side information. The quality of the side information has a major impact on the DVC Rate-Distortion (RD) performance in the same way the quality of the predictions had a major impact in predictive video coding. In this paper, a DVC solution exploiting multiple side information is proposed; the multiple side information is generated by frame interpolation and frame extrapolation targeting to improve the side information corresponding to a single estimation mode. Compared with the best available single side information solutions, the proposed DVC solution with multiple side information robustly improves the RD performance for the set of test sequences.

Index Terms— Distributed Video Coding, multiple side information, soft input.

1. INTRODUCTION

Distributed Video Coding (DVC) [1] proposes to fully or partly exploit the video redundancy at the decoder and not anymore at the encoder as in predictive video coding. According to the Slepian-Wolf theorem [2], it is possible to achieve the same rate by independently encoding but jointly decoding two statistically dependent signals as for typical joint encoding and decoding (with a vanishing error probability). The Wyner-Ziv theorem [3] extends the Slepian-Wolf theorem to the lossy case, becoming the key theoretical basis for Wyner-Ziv (WZ) video coding where some source is lossy coded based on the availability of some correlated source at the decoder from which the so-called side information is derived.

Feedback channel based transform domain Wyner-Ziv video codecs [4] are the most popular approaches to WZ video coding. Since the quality of the side information has a major impact on the final RD performance, there are several side information generation schemes proposed in the literature, notably frame interpolation [5] and frame extrapolation [6] based algorithms. Frame interpolation methods use previous and future decoded frames to generate the side information introducing some delay, while the extrapolation methods only use previously decoded frames. Generally, WZ coding with interpolated side information has better RD performance, notably for small GOP (Group of Pictures) sizes [6]. However, extrapolated side information has benefits for real-time applications due to the lower delay.

Since neither the available interpolation nor the extrapolation solution is perfect in terms of the created side information which is taken as estimation for the frames to WZ encode, the coding efficiency of Wyner-Ziv (WZ) video coding with single side information can be improved. The objective of this paper is to further progress the RD performance of WZ video coding, also reducing the RD gap regarding conventional video coding such as the H.264/AVC standard, by exploiting not a single but multiple side information. A first development in this area has been proposed in [7], where two different frame interpolation methods to generate the multiple side information are used. The channel decoder is fed with the average of two soft inputs which are generated based on two different side information estimates and the corresponding noise models. A more accurate soft input is obtained and the RD performance is improved up to 0.3 dB.

Differently, in this paper, the multiple side information is generated by frame interpolation and extrapolation. The intuition here is that having more different side information solutions should allow these to compensate each other's estimation weaknesses depending on the video content, overall leading to a more efficient coding solution. In this context, the extrapolated and the interpolated side information frames can be seen as original frames transmitted through quite different 'channels' and thus each side information frame is seen as an observation with a different amount of 'correlation noise'. With multiple observations, the WZ video decoder can select or combine the available side information estimations to decrease the amount of 'correlation noise' and thus to reduce misleading soft inputs in comparison with the single side information solution. In this way, the novel proposed solution shall reduce the required parity rate for each target quality, improving the RD performance.

The rest of this paper is organized as follows: Section 2 briefly describes the state-of-art on transform domain WZ video coding with feedback channel. In Section 3, the novel WZ decoder with interpolated and extrapolated side information is proposed. Finally, the test conditions and performance results are presented in Section 4.

2. STATE-OF-ART ON TRANSFORM DOMAIN WYNER-ZIV VIDEO CODING

A fixed Group of Pictures (GOP=*N*) is adopted in the state-of-art transform domain WZ video codec with feedback channel [4]. Periodically one frame out of *N* in the video sequence is named as key frame and intermediate frames are WZ frames. The key frames are intra coded by using a conventional video coding solution with low complexity such as H.264/AVC intra while the WZ frames are coded using a Wyner-Ziv video coding approach.

At the encoder, the WZ frames are partitioned into non-overlapped 4×4 blocks and an integer discrete cosine transform (DCT) is applied to each of them. The transform coefficients are grouped together and then quantized. After quantization, the coefficients are binarized, and each bitplane is given to a rate compatible Low Density Parity Check (LDPC) accumulate encoder [8] starting from the most significant bitplane. For each encoded bitplane, the corresponding accumulated syndrome is stored in a buffer at the encoder together with an 8-bit Cyclic Redundancy Check (CRC). The amount of bits to be transmitted depends on the requests made by the decoder through a feedback channel (Fig. 1).

The WZ decoder generates a side information frame Y by frame interpolation or extrapolation using previously decoded frames

[5][6]. Together with an estimated noise residue frame R, Y undergoes the integer DCT to obtain the coefficients C_Y and C_R. C_R is used to model the noise distribution between the corresponding DCT bands of the side information frame and the original WZ frame. Using the noise model [9], the coefficient values of the side information frame C_Y and the previous successfully decoded bitplanes, soft-input P (conditional bit probabilities) for each bitplane is estimated. With this soft-input P, the LDPC decoder starts to process the various bitplanes to correct the bit estimation errors. Convergence is tested by the 8-bit CRC sum and the Hamming distance between the received syndrome and the one obtained from the decoded bitplane: If the Hamming distance is different from zero or the CRC sum is incorrect after a certain amount of iterations, the LDPC decoder requests more accumulated syndrome bits from the encoder buffer via the feedback channel to correct the existing bit errors. If both the Hamming distance and CRC sum are satisfied, convergence is declared, guaranteeing a very low error probability for the decoded bitplane. For more details please refer to [4].

3. WYNER-ZIV DECODER WITH MULTIPLE SIDE INFORMATION

As mentioned before, the choice of the adopted side information generation scheme significantly influences the final coding efficiency. There are several interpolation and extrapolation methods in the literature, all targeting the generation of good quality side information frames [5][6]. The obtained side information frames are going to be used to estimate the soft-input information (conditional bit probabilities) for each bitplane based on a certain noise model [9]. The essential factor to reduce the number of coding bits is the soft-input information which is fed into the LDPC decoder. The more accurate the soft input is, the fewer parity bits are required by the decoder since the faster the convergence will be. Thus, an important way to increase the RD performance is to improve the soft-input information fed into the LDPC decoder.

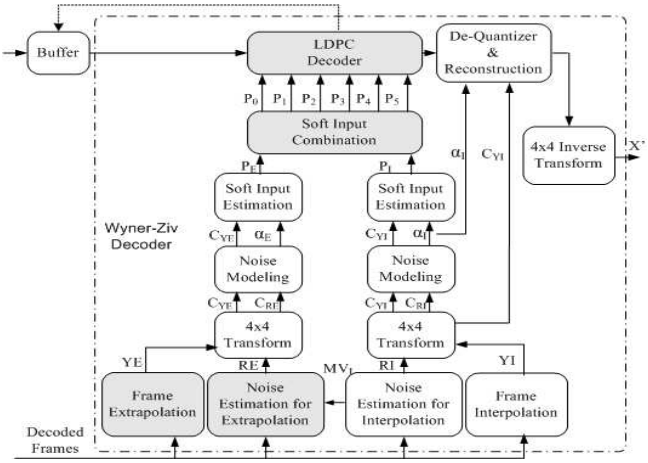

Fig. 1. *Transform domain Wyner-Ziv video decoder with interpolated and extrapolated side information*

The novel proposed WZ video codec with multiple side information follows this approach with the motivation described in Section 2. The encoder is not changed, as the basic idea is to generate better soft-input information by generating first better quality side information, in this case multiple side information through interpolation and extrapolation. While interpolation solutions are the most common in the literature, the WZ video codec proposed in this paper expects to improve the overall RD performance by also processing extrapolation side information which may be 'better' than interpolation side information for some conditions of the content. The architecture proposed for the novel WZ decoder with multiple side information is presented in Fig. 1. The track at the right starting with interpolation (*RI* and *YI*) presents a state-of-art WZ solution with interpolation. The technical novelty of the proposed WZ video decoder includes: i) an improved extrapolation method, ii) the noise estimation for extrapolation, iii) the soft inputs combination module, and iv) modified LDPC decoder.

3.1. WZ Decoder with Multiple Side Information Architecture

The main modules in the novel proposed WZ video decoder are:

• ***Frame Interpolation***: The adopted frame interpolation procedure is the same as in [5]. Without loss of generality, it generates the side information frame YI_{2i} by using intra coded frames, X'_{2i-1} and X'_{2i+1} for GOP size 2. It includes forward motion estimation, bi-directional motion estimation, spatial smoothing of Motion Vectors(MV), motion refinement with variable block size and adaptive weighted Overlapped Block Motion Compensation (OBMC). For more details, please refer to [5].

• ***Noise Estimation for Interpolation***: A motion estimated residue frame R_{ME} (i.e. the difference between X'_{2i-1} and X'_{2i+1} after motion compensation) is taken as the estimated noise residue *RI* to express the correlation noise between the WZ frame and the corresponding interpolated frame.

• ***Frame Extrapolation***: This module creates the extrapolated side information. The procedure is similar to [6]. Without loss of generality, the previous coded frames X'_{2i-1} and X'_{2i-2} are used to generate the side information frame YE_{2i} for GOP size 2. It includes motion estimation, spatial smoothing, frame projection, overlapping and filling holes. The difference is that a novel hole filling technique is applied. For the unreferenced/unfilled pixel areas in frame YE_{2i}, both the nearest MVs in the spatial domain and co-located MVs in temporal domain are used to determine the estimated pixels; an average of these estimates is computed for filling the holes remaining after the frame projection process.

• ***Noise Estimation for Extrapolation***: The noise residue *RE* is computed to present the correlation noise between the WZ frame and the corresponding extrapolated frame as described in Section 3.2.

• ***Noise Modeling***: After computing the 4×4 integer DCT coefficients C_{YI}, C_{YE}, C_{RI} and C_{RE} for the interpolated and extrapolated side information and the associated residues, the noise distribution between the side information and the corresponding WZ frames is estimated using a Laplacian noise model as described in [9]. Within a given DCT band b_k, the DCT coefficient at coordinates (m, n) is associated to the Laplacian parameter $\alpha_E^{bk}(m, n)$ for extrapolation and $\alpha_I^{bk}(m, n)$ for interpolation. The Laplacian parameter values express the reliability of the side information, i.e. the smaller this value is, the noisier the corresponding coefficient is.

• ***Soft Input Estimation***: With the obtained Laplacian parameters, side information coefficient values and the previous successfully decoded bitplanes, the soft-input information (conditional bit probabilities for extrapolation P_E and for interpolation P_I) of each bitplane are estimated [4].

• ***Soft Input Combination***: The soft input data to be provided to the LDPC decoder is generated by combining the soft inputs P_E and P_I in a few predefined modes creating various soft input candidates; see details in Section 3.3.

• ***LDPC Decoder***: All these candidate soft inputs are fed to a modified LDPC decoder. The soft input which converges (as described in Section 2) first is chosen by the LDPC decoder (Section 3.3) thus minimizing the rate of parity bits for a certain target quality.

• ***Reconstruction***: Based on the decoded bins, this module has to recover the coefficient's values also exploiting the available side information. Since the interpolated side information is typically better (see Fig. 2), the interpolated side information and its noise modeling

parameters are used by the reconstruction module [7] to recover the decoded WZ frames.

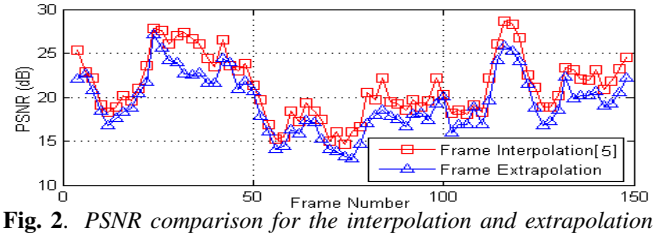

Fig. 2. *PSNR comparison for the interpolation and extrapolation methods for Soccer@15Hz, QCIF, GOP 2, Key frame H.264/AVC Intra coded, QP=25.*

3.2. Noise Estimation for Extrapolation

There are two natural ways to estimate the residue between WZ frames and the corresponding extrapolated side information to represent the correlation noise behavior:

• *Motion Estimated Residue* R_{ME}: Corresponds to the pixel differences between X'_{2i-1} and X'_{2i-2} along the extrapolated MVs.

• *No Motion Estimated Residue* R_{NO}: Corresponds to the co-located pixel differences between YE_{2i} and X'_{2i-1}.

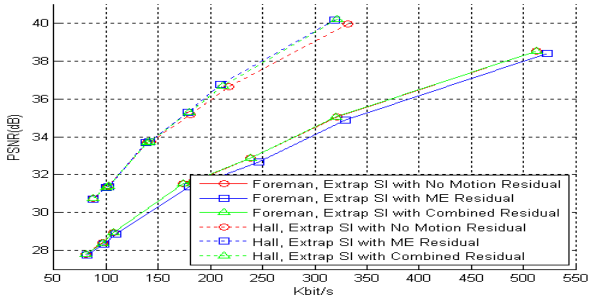

Fig. 3. *RD performances with extrapolated side information using the motion estimated and no motion estimated residues for Foreman and Hall Monitor, QCIF, 15 Hz.*

Experiments have shown that, when creating the side information using frame extrapolation, the more commonly used motion estimated residue [9] will provide a worse RD performance for high motion sequences while it will perform better for low motion sequences in comparison with the no motion estimated residue (see Fig. 3). The worse RD performance may be caused by the linear motion assumption adopted for the generation of the unidirectional MVs used for the frame extrapolation process. If these MVs are not fulfilling this assumption, then the extrapolated block is going to be projected into a wrong position, corresponding to a large real noise residue, while the motion estimated residue R_{ME} will be smaller. Based on this poorly estimated noise residue, the estimated Laplacian parameter will be inaccurate in terms of noise modeling, misleading the LPDC decoder in terms of the soft input P_E. In order to solve this problem, it is necessary to generate a more robust estimate for the noise residue when frame extrapolation is used. In this context, it is proposed here to check the 'accuracy' of the motion vectors obtained by extrapolation MV_E using the motion vectors obtained by frame interpolation MV_I. The intuition is that if the two sets of MVs are similar, then the motion description should be good and thus the motion estimated residue should be used. Following this intuition, a combined noise residue, R_{COM}, is computed by switching between R_{ME} and R_{NO} as:

$$R_{COM}(x,y) = \begin{cases} R_{ME}(x,y), & \text{if } MV_I(m,n)=MV_E(m,n) \\ R_{NO}(x,y), & \text{otherwise} \end{cases} \quad (1)$$

where (x, y) are the pixel coordinates and (m, n) are the corresponding block coordinates. The RD performance with single extrapolation side information using the proposed combined noise residue is compared with the relevant alternatives in Fig. 3 for the Foreman and Hall Monitor sequences.

3.3. Soft Input Combination

After the extrapolation soft input P_E and the interpolation soft input P_I are obtained, the soft input combination module has the task of adaptively combining these two soft inputs to generate a set of candidate soft inputs, thus improving the RD performance by reducing the rate of parity bits.

Since the values of the Laplacian parameters should express the reliability of the corresponding side information, an unreliability region map is defined as the region of the frame where extrapolation or interpolation indicates areas including discontinuous linear motion. It means there should be little benefit brought by extrapolation outside of the map region within which the motion is relative linear. This map region is determined by evaluating the Laplacian parameters and their corresponding mean value as:

$$map = \{(m,n) | \alpha_E^{bk}(m,n) < E(\alpha_E^{bk}) \vee \alpha_I^{bk}(m,n) < E(\alpha_I^{bk})\} \quad (2)$$

where $\alpha_E^{bk}(m,n)$ and $\alpha_I^{bk}(m,n)$ are the estimated Laplacian distribution parameters within DCT band b_k for extrapolation and interpolation, respectively. (m, n) are the block coordinates and $E(\alpha^{bk})$ represents the mean value of the Laplacian parameter over all the blocks within DCT band b_k.

In order to take advantage of the benefits brought by the extrapolation soft input P_E regarding a single interpolation side information solution, a set of candidate soft inputs is generated by combining the extrapolation soft input P_E with the interpolation soft input P_I within the unreliability region map, while only the interpolation soft input P_I is adopted in the reliable region (there is no expected benefit in also using P_E):

$$P_T(m,n) = \begin{cases} w_T \cdot P_I(m,n) + (1-w_T) \cdot P_E(m,n), & \\ \qquad \text{if } (m,n) \in map & \\ P_I(m,n), \quad \text{otherwise} & \end{cases} \quad (3)$$

where $w_T = \{1 - (T/10) | T = 0, 1, 2, 3, 4, 5\}$. All these candidate soft inputs are fed into the LDPC decoder; the one which first converges will be chosen thus reducing the rate of parity bits for the same target quality. By using this set of combined soft inputs, the extrapolation side information track will influence the LDPC decoding process, reducing the amount of misleading soft inputs provided by the interpolation side information track, following the intuition behind this paper and reaching the stated objective of improving the overall RD performance based on more and better side information.

4. EXPERIMENTAL RESULTS

In order to make fair comparisons, the test conditions adopted in this paper are the DISCOVER project test conditions, commonly used in the DVC literature [4]. The test sequences are *Foreman*, *Soccer*, *Coastguard* and *Hall Monitor*, coded at QCIF, 15 frames per second (fps); the GOP size is 2. The key frames are encoded using H.264/AVC Intra and the QPs are chosen so that the average PSNR of the WZ frames is similar to the average PSNR of the key frames (as in [4]). The RD performance is evaluated for the luminance component of both the key frames and WZ frames. The benchmark codecs used are the DISCOVER WZ video codec [4] and the H.264/AVC Intra codec. For comparison, the performance of some other relevant transform domain WZ video codecs with single (interpolation [5] or extrapolation) and multiple (interpolation and extrapolation) side information is also included.

As shown in Figs. 4-7, the performance of the single interpolation side information WZ video codec is better than the DISCOVER codec due to the OBMC based interpolation side information method [5]. The RD performance with single interpolation side information is better than the one with single extrapolation side information meaning that the additional delay involved really brings additional RD performance. Moreover, based on precisely the same

H.264/AVC intra coded key frames, the multiple side information codec can improve the overall RD performance of single interpolation side information codec up to 0.4 dB at high bitrates for the WZ frames. Since the interpolation side information is quite efficient for low motion sequences, the extrapolation side information brings less RD performance improvements in the context of WZ coding with multiple side information for this type of video content. This means that compared with low motion sequences like *Hall Monitor*, WZ decoding with multiple side information provides larger RD gains for high motion sequences like *Foreman* and *Soccer*. WZ video coding with multiple side information already gives better RD performance than H.264/AVC intra coding for *Foreman*, *Coastguard* and *Hall Monitor*; for sequences with more irregular motion like *Soccer*, where the decoder frame estimation process is more difficult, the performance gap between H.264/AVC intra coding and WZ video coding has been reduced but not yet closed.

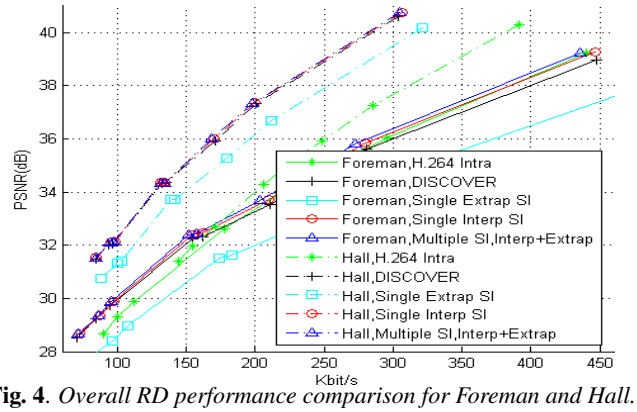

Fig. 4. *Overall RD performance comparison for Foreman and Hall.*

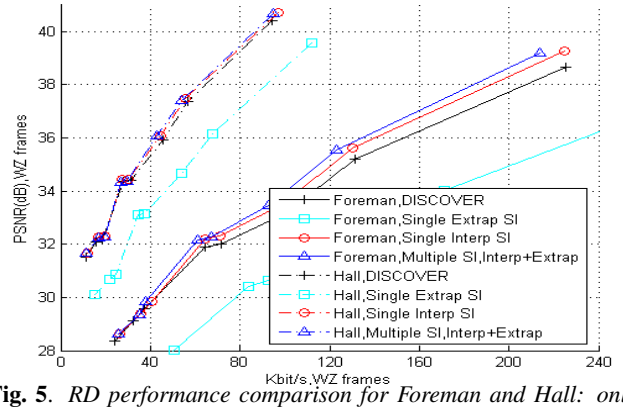

Fig. 5. *RD performance comparison for Foreman and Hall: only WZ frames for precisely the same key frames.*

5. CONCLUSION

A novel transform domain WZ video decoder with multiple (interpolation and extrapolation) side information is proposed in this paper with the objective to improve the overall RD performance. Although the extrapolated side information frames are significantly worse than the interpolated side information frames, improvement is robustly achieved by generating and combining a set of candidate soft inputs to be fed to the LDPC decoder, trying to reduce the number of bits requested by the decoder for a target quality; this process implies adaptively to combine the interpolation and extrapolation derived soft inputs with the aim of using the most reliable side information derived soft input depending on the video content. Compared with state-of-art single side information WZ video coding solutions, the proposed transform domain WZ video codec with multiple side information can improve the overall RD performance for the set of test sequences; the RD gains may go up to 0.4 dB (averaged over the sequence) for the WZ frames with precisely the same H.264/AVC intra coded key frames.

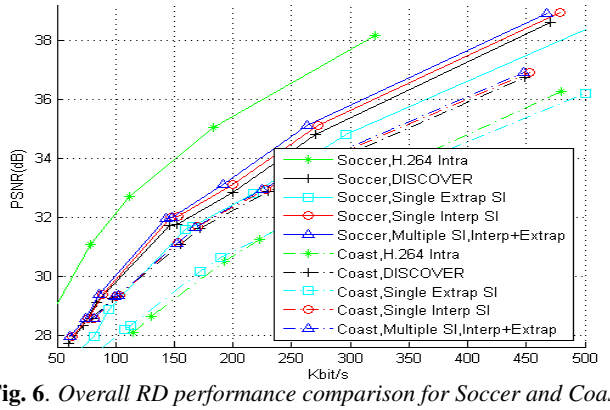

Fig. 6. *Overall RD performance comparison for Soccer and Coast.*

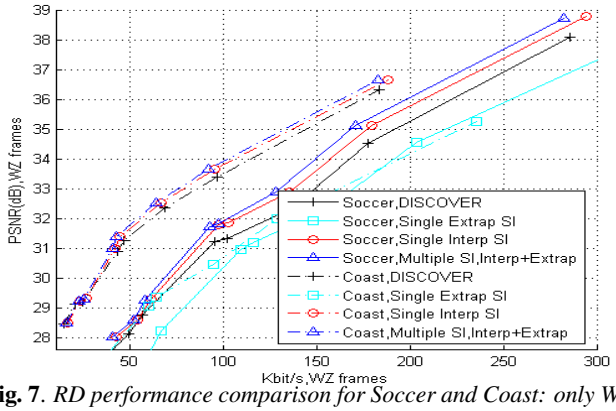

Fig. 7. *RD performance comparison for Soccer and Coast: only WZ frames for precisely the same key frames.*

6. REFERENCES

[1] A. Aaron, S. Rane, E. Setton, and B. Girod, "Transform domain wyner-ziv codec for video," *Proc. SPIE VCIP*, San Jose, USA, Jan. 2004.

[2] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. 19, no.4, pp. 471–480, July 1973.

[3] A.D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. 22, no.1, pp. 1–10, Jan. 1976.

[4] DISCOVER Project: www.discoverdvc.org, Dec. 2007.

[5] X. Huang and S. Forchhammer, "Improved side information generation for distributed video coding," *IEEE Int'l Workshop Multimedia Signal Process.*, Cairns, Australia, Oct. 2008.

[6] L. Natário, C. Brites, J. Ascenso, and F. Pereira, "Extrapolating side information for low-delay pixel-domain distributed video coding," *Int'l Workshop on Very Low Bitrate Video Coding*, Sardinia, Italy, Sept. 2007.

[7] D. Kubasov, J. Nayak, and C. Guillemot, "Optimal reconstruction in wyner-ziv video coding with multiple side information," *IEEE Int'l Workshop Multimedia Signal Process.*, Chania, Greece, Oct. 2007.

[8] D. Varodayan, A. Aaron, and B. Girod, "Rate-adaptive distributed source coding using low-density parity-check codes," *EURASIP Signal Process. Journal, Special Section on Distributed Source Coding*, vol. 86, pp. 3123–3130, Nov. 2006.

[9] C. Brites and F. Pereira, "Correlation noise modelling for efficient pixel and transform domain wyner-ziv video coding," *IEEE Trans. on Circuits. Syst. Video Technol.*, vol. 18, no.9, pp. 1177–1190, Sept. 2008.