SUCCESSIVE REFINEMENT OF MOTION COMPENSATED INTERPOLATION FOR TRANSFORM-DOMAIN DISTRIBUTED VIDEO CODING

A. Abou-Elailah [1], *F. Dufaux* [1], *J. Farah* [2], *M. Cagnazzo* [1], *and B. Pesquet-Popescu* [1]

[1] Signal and Image processing Department, Institut TELECOM - TELECOM Paristech,
46 rue Barrault, F - 75634 Paris Cedex 13, FRANCE
{elailah, frederic.dufaux, marco.cagnazzo, beatrice.pesquet}@telecom-paristech.fr
[2] Engineering Department, Faculty of Engineering, Holy-Spirit University of Kaslik
P.O. Box 446, Jounieh, Lebanon
joumanafarah@usek.edu.lb

ABSTRACT

In distributed video coding, the estimation of the side information at the decoder plays a key role in the final rate-distortion performance of the codec. The side information is commonly generated by motion-compensated temporal interpolation of the neighboring reference frames. In this paper, we propose a successive refinement after the decoding of each DCT subband to improve the accuracy of motion compensation between reference frames, in order to obtain a new side information estimation closer to the original Wyner-Ziv frame. The experimental results show that the proposed scheme can achieve up to 0.9 dB of improvement in rate-distortion performance for a GOP size of 2 and 2.4 dB for a GOP size of 8 for sequences containing high motion with respect to state-of-the-art techniques.

1. INTRODUCTION

Nowadays, many digital video coding solutions are available, such as the ISO/IEC MPEG and ITU-T H.26x standards [1] based on DCT, inter-frame and intra-frame predictive coding. In these standards, the encoder is significantly more complex than the decoder due to the exploitation of the video redundancies at the encoder. This kind of architecture is well-suited for applications where the video sequence is encoded once and decoded many times, such as broadcasting or video streaming. However, this architecture is being challenged by several emerging applications such as wireless video surveillance, multimedia sensor networks, wireless PC cameras, and mobile camera phones. In these cases, new requirements arise such as low complexity encoding.

Distributed Video Coding (DVC) fits well these emerging scenarios since it enables to exploit video redundancies at the decoder side only, while the encoder is less complex. In other words, the complexity is shifted from the encoder to the decoder. From information theory, the Slepian-Wolf theorem for lossless compression [2] states that it is possible to encode correlated sources (let us call them X and Y) independently and decode them jointly, using a rate similar to that used in a system where sources are encoded and decoded jointly. The Wyner-Ziv theorem [3] extends the Slepian-Wolf theorem to lossy compression, which deals with lossy source coding of X when Side Information (SI) Y is available at the decoder only.

In recent years, practical implementations of DVC [4, 5] have been proposed based on these theoretical results. We consider here the DISCOVER codec [6, 7] which is based on transform domain Wyner-Ziv coding. In this codec, the video sequence is splitted into two sets of frames: the Key Frames (KFs), and the Wyner-Ziv Frames (WZFs). KFs are encoded without using temporal predictions (i.e. an Intra coding technique is used such as JPEG2000 or the H.264/AVC Intra mode). At the decoder, the decoded KFs are used in order to compute the SI, which is an estimation of the WZF being decoded. The technique used in the DISCOVER codec to generate the SI is based on Motion-Compensated Temporal Interpolation (MCTI) [8].

SI quality has a strong impact on the final Rate-Distortion (RD) performance. For this reason, many approaches have been proposed in order to improve the SI quality at the decoder. For example, a method proposed by Aaron et al. [9] and by Ascenso et al. [10] consists in sending a hash of WZF information to enhance the interpolation of the SI. However, these techniques demand some additional data (the hash) to be sent through the channel. Other techniques exist which can avoid this overhead. They are based on the successive refinement of the SI. A solution proposed by J. Ascenso et al. [11] for pixel domain DVC uses a motion compensated refinement of the SI successively after each decoded bit plane in order to achieve a better reconstruction of the decoded WZF. In [12], the authors proposed a novel DVC successive refinement approach to improve the motion compensation accuracy and the SI. This approach is based on the N-Queen sub-sampling pattern. The authors in [13] proposed a solution for transform-domain DVC, which refines the SI after the decoding of all DCT bands in order to improve the reconstruction.

In [14, 15], solutions are proposed for transform-domain DVC based on the successive refinement of the SI after each decoded DCT band. The authors in [15] only used the SI and PDWZ to refine the SI for decoding the next band, although there is more information to be extracted from the backward and forward reference frames. The solution proposed in [14] relies on a much simpler SI refinement technique compared with our proposed approach. Moreover, the refinement procedure is applied to all blocks regardless of the motion vector reliability.

In this paper, we propose an approach in order to enhance the SI in transform-domain DVC. This solution consists in progressively improving the SI after each decoded DCT-band and is particularly efficient for high motion regions and in the case where KFs are separated by a significant number of WZFs. We first start by generating an Initial Side Information (INSI) by using the backward and forward reference

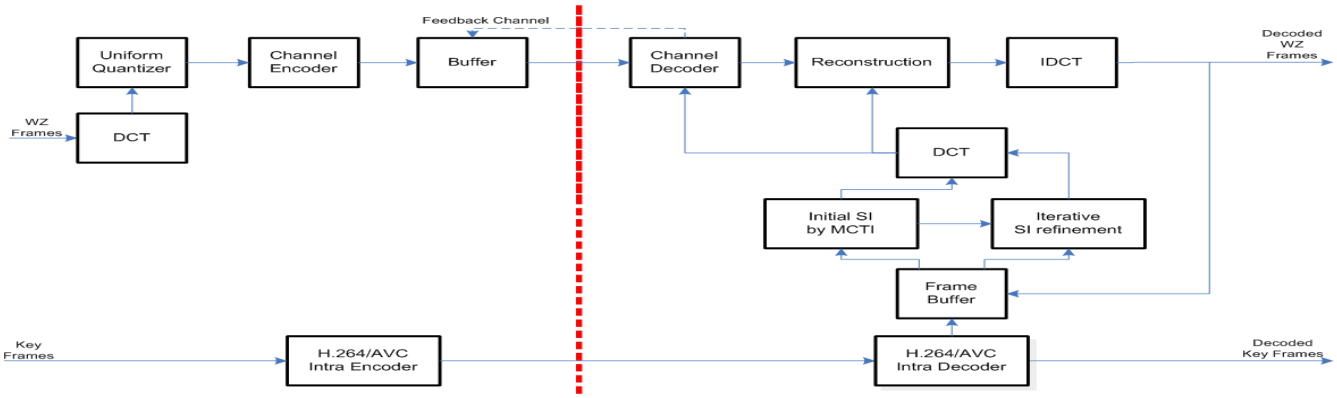


Figure 1: Overall structure of proposed DVC coding.

frames similarly to the SI generated in DISCOVER. The decoder reconstructs a Partially Decoded Wyner-Ziv (PDWZ) frame by correcting the INSI with the parity bits of the first DCT-band. Then, the PDWZ frame along with the backward and forward reference frames is used to refine the INSI. The refinement method consists of four modules: Suspicious Vector Detection, Refinement, Mode Selection, and Motion Compensation (see section 2.2 for more details). Finally, we correct this refined INSI with the parity bits of the next DCT-band, and we repeat the same procedure to decode all DCT-bands of the current WZF.

This paper is structured as follows. First, the proposed system architecture based on the DISCOVER codec and the proposed approach by successive refinement of the SI are described in Section 2. Experimental results are then showed in Section 3 in order to evaluate and compare the RD performance of the proposed approach. Finally, conclusions and future work are presented in Section 4.

2. PROPOSED SYSTEM ARCHITECTURE

The block diagram of our proposed codec architecture is depicted in Figure 1. It is based on the DISCOVER codec [6, 7]. First, the video sequence is divided into WZFs and KFs. The Group of Pictures (GOP) is defined as the distance between two consecutive KFs which are coded using H.264/AVC Intra coding. The WZF encoding and decoding procedures are detailed in the following. We start by briefly recalling the DISCOVER architecture, and we describe the proposed scheme by difference.

2.1 DISCOVER codec modules

- Wyner-Ziv encoder - At the encoder side, the WZF is first transformed using a 4×4 block-based Discrete Cosine Transform (DCT). The DCT coefficients of the entire WZF are then organized in 16 bands, indicated by b_k with $k \in [1,16]$, according to their position within the 4×4 blocks. The low frequency information (i.e. the DC coefficients) are placed in the first band $k = 1$, and the other coefficients are grouped in the AC bands $k = 2,3,...16$. Next, each DCT coefficients band b_k is uniformly quantized with 2^{M_k} levels. The resulting quantized symbols are then split into bit planes. For a given band, the quantized symbols bits of the same significance are grouped together in order to form the corresponding biplane array which is then independently encoded using a channel encoder. The latter, also known as the Slepian-Wolf encoder, is a rate-compatible Low-Density Parity Check (LDPC) accumulate code. Each bitplane is successively fed into the channel encoder in order to compute a separate set of parity bits, while the systematic bits are discarded. The parity information is then stored in a buffer and progressively sent in chunks, upon request by the decoder, through the feed-back channel.
- Generation of side information - In the DISCOVER scheme, the frame interpolation framework is composed of four modules to obtain high quality SI [8] (preceded by low-pass filtering of the reference frames in order to improve the motion vectors reliability): forward motion estimation between the previous and next reference frames, bi-directional motion estimation to refine the motion vectors, spatial smoothing of motion vectors in order to achieve higher motion field spatial coherence (reduction of the number of false motion vectors), and finally bi-directional motion compensation.
- Wyner-Ziv decoder - A block-based 4×4 DCT is carried out over the generated SI (using the previous step) in order to obtain the DCT coefficients which can be seen as a noisy version of the WZF DCT coefficients. In order to model the error distribution between corresponding DCT bands of SI and WZF, the DISCOVER codec uses a Laplacian distribution. The Laplacian parameter is estimated on-line at the decoder. Once the DCT transformed SI and the residual statistics for a given DCT band b_k are known, the Slepian-Wolf decoder corrects the bit errors in the DCT transformed SI using the parity bits of WZF requested through the feedback channel.
- Reconstruction and inverse transform - The reconstruction corresponds to the inverse of the quantization using the SI DCT coefficients and the decoded WZF DCT coefficients. After that, the inverse 4×4 DCT transform is carried out, and the entire frame is restored in the pixel domain.

2.2 Proposed modules - Successive Refinement and Mode Selection

The INSI is first computed by MCTI with spatial motion smoothing exactly as in DISCOVER codec. The LDPC parity bits of the first band (DC band) are used in order to correct the correspondent DCT coefficients in INSI. The obtained

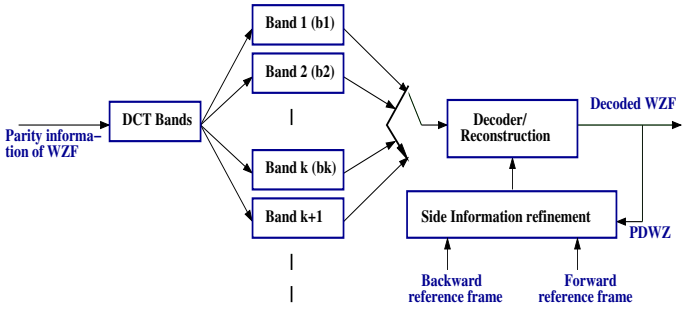


Figure 2: Proposed technique for successive refinement of SI and WZF decoding.

Foreman - GOP size = 2					
	$T_2 = 4$				
	$T_1 = 0$	$T_1 = 2$	$T_1 = 4$	$T_1 = 6$	$T_1 = 8$
Δ_R (%)	-11.48	-13.22	**-13.40**	-11.00	-8.98
Δ_{PSNR} [dB]	0.60	0.71	**00.73**	0.60	0.47
	$T_1 = 4$				
	$T_2 = 0$	$T_2 = 2$	$T_2 = 4$	$T_2 = 6$	$T_2 = 8$
Δ_R (%)	-9.88	**-13.07**	**-13.40**	**-13.41**	**-13.43**
Δ_{PSNR} [dB]	0.54	**0.70**	**0.73**	**0.73**	**0.73**

Table 1: RD performance gain for *Foreman* for different values of T_1 and T_2, towards DISCOVER codec, using Bjontegaard metric.

decoded frame is denoted as Partially Decoded Wyner-Ziv (PDWZ) frame. Here, we use the two adjacent reference frames and the PDWZ in order to improve the SI interpolation. The proposed approach is similar to the method in [13]. However, it has been improved in such a way that the SI is progressively refined after the decoding of each DCT-band. Moreover, both the matching criterion and the mode selection have been modified, resulting in improved performances.

The proposed scheme for SI enhancement consists of three steps based on PDWZ to improve the SI: suspicious vector detection based on the matching criterion, motion vector refinement and smoothing, and motion compensation mode selection.

- Matching criterion - In order to exploit the temporal correlations to enhance the estimated motion vectors, the matching criterion used in this paper is based on the Mean Absolute Difference (MAD). The MAD for the estimated motion vector MV of a Block B is defined as :

$$\begin{aligned} &\mathrm{MAD}(F_c, F_r, \mathbf{MV}) = \tfrac{1}{M\times N}\textstyle\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} \\ &|F_c(i,j) - F_r(i+MV_x, j+MV_y)| \end{aligned} \quad (1)$$

  where F_c is the current frame, F_r the reference one, and $\mathbf{MV} = (MV_x, MV_y)$ the candidate motion vector. The current block has M rows and N columns. In this paper, the size of the block is considered to be 8×8 pixels as in DISCOVER.

- Suspicious vector detection - The motion vectors estimated by MCTI for sequences with low motion are close to the true motion. However, false motion vectors may occur in sequences with high motion and occlusions. In order to identify suspicious vectors, a threshold T1 is used. We estimate the MAD for a given block between the PDWZ and the last SI (e.g. previous refinement of the SI):

$$\mathrm{MAD}(F1, F2(\mathbf{MVB}), \mathbf{0}) < T_1, \quad (2)$$

  where $\mathbf{0} = (0,0)$ is the null vector, $F1$ and $F2$ are the PDWZ frame and the last refinement SI respectively. If this estimation satisfies the condition defined in equation (2), it is considered to be a true estimation (e.g. the motion vector $(\mathbf{MVB})$ for this block is not modified). Otherwise, it is identified as a suspicious vector and will be further refined.

- Motion vector refinement - In order to refine the motion vectors which are identified as suspicious vectors, we re-estimate these motion vectors by bi-directional motion estimation using the matching criterion defined in equation (1). For the current block in the PDWZ frame, we look for the motion vector which minimizes the MAD within a window in the previous reference frame. The motion vector between the PDWZ and the forward reference frame is estimated in the same way. These estimated motion vectors are considered as refined motion vectors for the processed block.

- Motion compensation mode selection - The objective of this step is to generate an optimal motion-compensated estimate by selecting the most similar block to the current block from three sources: the previous reference frame (BACKWARD MODE), the next reference frame (FORWARD MODE), and bi-directional motion-compensated average of the previous and next reference frames (BIMODE). The decision among these modes is performed according to the following equations, and a threshold T_2 is established.

$$\begin{cases} \text{if } |\mathrm{MAD}_n - \mathrm{MAD}_p| < T_2 \\ \quad \text{MODE=BIMODE} \\ \text{otherwise} \\ \quad \text{if } \mathrm{MAD}_n < \mathrm{MAD}_p \\ \quad\quad \text{MODE=FORWARD MODE} \\ \quad \text{otherwise} \\ \quad\quad \text{MODE=BACKWARD MODE} \end{cases}$$

  where MAD_p and MAD_n are the estimated mean absolute differences between the current block (in PDWZ) and the corresponding blocks (e.g. the blocks which minimize the MAD) in the previous and next reference frames respectively.

The refinement SI obtained (after decoding the band b_k) is used at the WZ decoder as a new SI for the decoding of the next band b_{k+1}, and so forth for all bands of the WZF being decoded. Then, after decoding all bands, a new SI is generated to perform again the reconstruction step to get the final WZF. The proposed scheme for this procedure is illustrated in Figure 2.

3. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed codec, we performed extensive simulations, adopting the same test conditions as described in DISCOVER codec [6, 7] (test video sequences are at QCIF spatial resolution and sampled at 15 frames/sec). Our obtained results are compared to simulation results of DISCOVER codec, the method in [13], H.264/AVC No motion, and H.264/AVC Intra.

Table 1 shows the RD performance gain for Foreman sequence for different values of T_1 and T_2, using Bjontegaard metric [16]. We have set $T_1 = 4$ due to the high performance

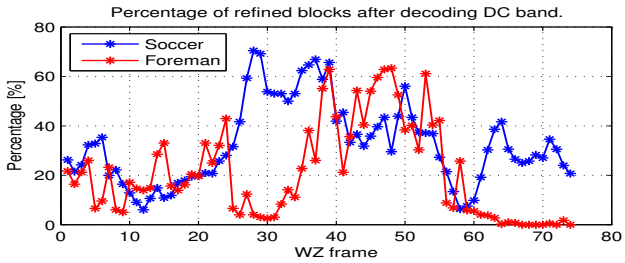
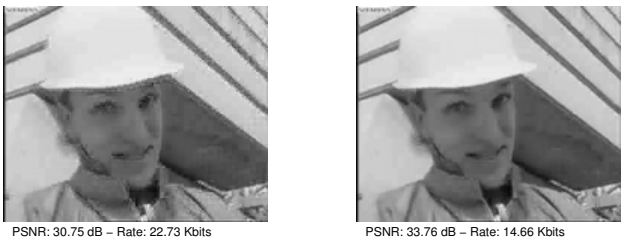


Figure 3: Percentage of refined blocks after decoding the DC band for Soccer and Foreman sequences with a GOP size 2.

(a) The decoded frame (number 78, GOP=2) - Soccer sequence.



(b) The decoded frame (number 76, GOP=2) - Foreman sequence.

Figure 4: Visual result comparisons between the proposed method (right) and DISCOVER codec (left).

and low computational load. It is clear that the bidirectional mode ($T_2 > 0$) is better than the unidirectional mode ($T_2 = 0$), we have set $T_2 = 4$.

Figure 3 indicates the percentage of refined blocks, that is, blocks which are identified as suspicious MV per execution of refinement after decoding the DC band, for Soccer and Foreman sequences, with a GOP size equal to 2. It is clear that the percentage of refined blocks increases with the motion within the video sequence. For Foreman sequence, the percentage tends to zero due to the low motion at the end of the sequence.

Figure 4 shows the visual results of the decoded frames for Soccer and Foreman sequences for a GOP equal to 2. The decoded frames obtained by DISCOVER codec contain block artifacts. On the contrary, the decoded frames obtained by the proposed method are better, up to 2 dB improvement for Soccer and 3 dB for Foreman, with less requested bits, down from 26.57 Kbits to 21.34 Kbits for Soccer and down from 22.73 Kbits to 14.66 Kbits for Forman.

The RD performance of the proposed method for the Soccer sequence is shown in Figure 5 with different GOP sizes (2, 4, and 8). RD performance of our scheme is better than the DISCOVER codec and [13] for all GOP sizes. Our proposed method allows a gain of 0.92 dB over the DISCOVER codec for GOP size equal to 2. The RD performance of our scheme is better than the DISCOVER codec by 1.45 dB for GOP length equal to 4, and improves significantly the objective quality (up to 1.65 dB) with respect to DISCOVER codec with GOP size equal to 8. With respect to [13], the gains are respectively 0.64, 0.95 and 1.08 dB for GOP size 2, 4 and 8. It is clear that the gain in RD performance increases with the GOP length. In this case, classical interpolation techniques for SI generation become less effective.

We also show the RD performance of our scheme for Foreman sequence with different GOP lengths (2, 4, and 8) in Figure 6. The RD performance of our method is, again, better than the DISCOVER codec for all GOP sizes. The proposed method achieves an improvement in RD performance up to 0.75 dB with GOP size equal to 2 compared to DISCOVER codec, and a significant improvement, up to 2.4 dB, in RD performance compared to DISCOVER codec with GOP length equal to 8. Compared to [13], the gains become respectively 0.5, 1.1 and 1.55 dB for GOP size 2, 4 and 8. For a GOP length equal to 4, our proposed method allows a gain of 1.7 dB over DISCOVER codec. For Hall Monitor and Coastguard sequences, the gains are smaller and range between 0.2 and 0.8 dB.

The difference between the performance of DISCOVER and H.264/AVC No motion and Intra is significant for Soccer, the proposed method can reduce this difference by half. For Foreman, the performance of the proposed system is better than the H.264/AVC Intra for all GOP sizes, and the difference between the proposed approach and H.264/AVC No motion becomes small.

It can be noticed that the performance gains are more substantial for high motion sequences and for long GOP sizes. They are mainly associated with the proposed technique for successive refinement of the side information interpolation, which is the major contribution with respect to the reference codec. The proposed method enhances the quality of the SI, as well as, the time of the decoding process is reduced due to the less demanding of parity bits via the feedback channel to correct the errors in the SI.

4. CONCLUSIONS

Successive refinement of the side information using the partially decoded frame and the two adjacent reference frames in DISCOVER codec was proposed in this paper, based on the successive decoding of the DCT bands. Experimental results showed that our proposed method can achieve a better RD performance compared to DISCOVER codec, especially when the video sequence contains high motion. The improvement becomes even more important as the GOP size increases.

5. ACKNOWLEDGEMENT

This work was partly supported by a research grant from the Franco-Lebanese CEDRE program and PERSEE project .

REFERENCES

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

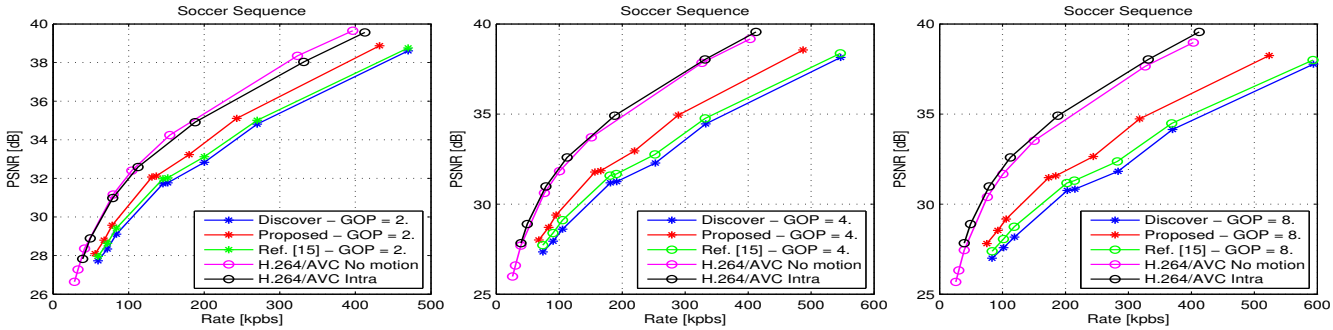


Figure 5: RD performance for Soccer sequence with GOP = 2, 4, and 8.

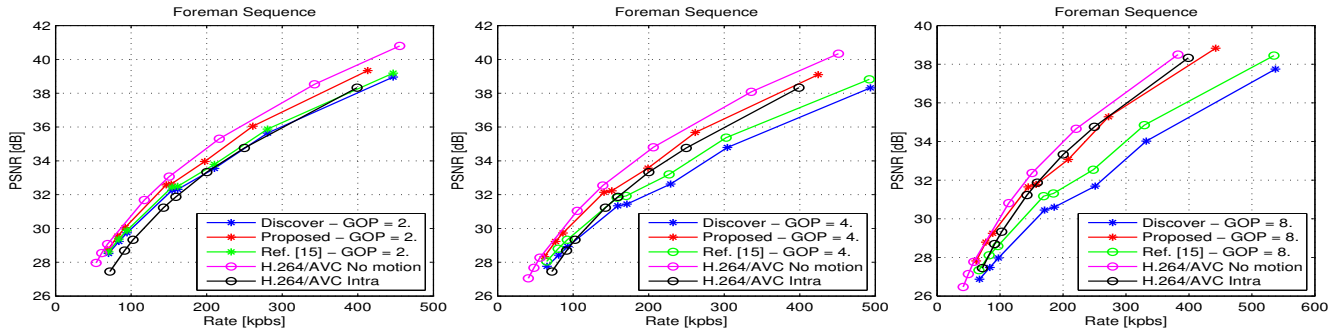


Figure 6: RD performance for Foreman sequence with GOP = 2, 4, and 8.

[2] J.D. Slepian and J.K. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. IT-19, pp. 471–480, July 1973.

[3] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, pp. 1–10, July 1976.

[4] R. Puri and K. Ramchandran, "PRISM: A video coding architecture based on distributed compression principles," *EECS Department, University of California, Berkeley, Tech. Rep. UCB/ERL M03/6*, 2003.

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

[6] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M.Ouaret, "The DISCOVER codec: Architecture, techniques and evaluation," *Proc. of Picture Coding Symposium*, Oct. 2007.

[7] "Discover project," `http://www.discoverdvc.org/`.

[8] C. Brites, J. Ascenso, and F. Pereira, "Improving transform domain Wyner-Ziv video coding performance," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, vol. 2, pp. 525–528, May 2006.

[9] A. Aaron, S. Rane, and B. Girod, "Wyner-Ziv video coding with hash-based motion compensation at the receiver," *Proc. Int. Conf. on Image Processing*, vol. 05, pp. 3097–3100, Oct. 2004.

[10] J. Ascenso and F. Pereira, "Adaptive hash-based side information exploitation for efficient Wyner-Ziv video coding," *Proc. Int. Conf. on Image Processing*, vol. 03, pp. 29–32, Oct. 2007.

[11] J. Ascenso, C. Brites, and F. Pereira, "Motion compensated refinement for low complexity pixel based distributed video coding," *Proceedings of the IEEE international conference on Advanced Video and Signal-Based Surveillance*, pp. 593 – 598, Sept. 2005.

[12] X. Fan, O. Au, N. Cheung, Y. Chen, and J. Zhou, "Successive refinement based Wyner-Ziv video compression," *Signal Processing: Image Communication*, vol. 25, pp. 47–63, Jan. 2010.

[13] S. Ye, M. Ouaret, F. Dufaux, and T. Ebrahimi, "Improved side information generation for distributed video coding by exploiting spatial and temporal correlations," *EURASIP Journal on Image and Video Processing*, vol. 2009, pp. 15 pages, 2009.

[14] M.B. Badem, W.A.C. Fernando, J.L. Martinez, and P. Cuenca, "An iterative side information refinement technique for transform domain distributed video coding," *IEEE International Conference on Multimedia and Expo, ICME*, pp. 177 – 180, 2009.

[15] R. Martins, C. Brites, J. Ascenso, and F. Pereira, "Refining side information for improved transform domain wyner-ziv video coding," *IEEE Transactions on circuits and systems for video technology*, vol. 19, no. 9, pp. 1327 – 1341, Sept. 2009.

[16] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," in *VCEG Meeting*, Austin, USA, Apr. 2001.