Refining WZ rate estimation in DVC with feedback channel constraints

Slowack, Jürgen, Lambert, Peter, Van de Walle, Rik

Jürgen Slowack, Peter Lambert, Rik Van de Walle, "Refining WZ rate estimation in DVC with feedback channel constraints," Proc. SPIE 8499, Applications of Digital Image Processing XXXV, 84990L (15 October 2012); doi: 10.1117/12.929536

SPIE.

Event: SPIE Optical Engineering + Applications, 2012, San Diego, California, United States

Refining WZ rate estimation in DVC with feedback channel constraints

Jürgen Slowack[a,b], Peter Lambert[a,b], and Rik Van de Walle[a,b]

[a]Ghent University, ELIS - Multimedia Lab, Gaston Crommenlaan 8 bus 201, B-9050 Ledeberg-Ghent, Belgium;
[b]IBBT, Gaston Crommenlaan 8 bus 102, B-9050 Ledeberg-Ghent, Belgium

ABSTRACT

Distributed video coding (DVC) has attracted a lot of attention during the past decade as a new solution for video compression where the computationally most intensive operations are performed by the decoder instead of by the encoder. One very important issue in many current DVC solutions is the use of a feedback channel from the decoder to the encoder for the purpose of determining the rate of the coded stream. The use of such a feedback channel is not only impractical in storage applications but even in streaming scenarios feedback-channel usage may result in intolerable delays due to the typically large number of requests for decoding one frame. Instead of reverting to a feedback-free solution by adding complexity to the encoder for performing encoder-side rate estimation, as an alternative, in previous work we proposed to incorporate constraints on feedback channel usage. To cope better with rate fluctuations caused by changing motion characteristics, in this paper we propose a refined approach exploiting information available from already decoded frames at other temporal layers. The results indicate significant improvements for all test sequences (using a GOP of length four).

Keywords: Distributed Video Coding, Wyner-Ziv coding, feedback channel

1. INTRODUCTION

Video compression is typically achieved by exploiting the spatial and temporal redundancies present in the video sequence. Conventional schemes (e.g., based on H.264/AVC or MPEG-4) exploit such redundancies at the encoder. Due to the complexity of these operations, conventional schemes feature a complexity imbalance with an encoder that is (often many times) more complex than the decoder. This fits very well down-link scenarios where a video is encoded offline or by high-end devices and decoded possibly many times by users with low-end devices (such as mobile phones and set-top boxes).

Distributed video coding (DVC) offers an alternative solution for video compression by exploiting the correlations in the sequence at the decoder, resulting in a complex decoder but relatively simple encoder. Such a complexity distribution fits better up-link scenarios where encoding devices are power-constrained, small, and/or cheap, fitting applications such as wireless video surveillance, wireless visual sensor networks, and wireless capsule endoscopy.[1] An additional advantage for DVC is that – in a multi-view context – no communication between the encoders is needed as the inter-view correlations can be exploited at the decoder only.[2] Also, DVC techniques can be used to provide or facilitate other features such as error resilience.[3]

In DVC, the basic idea at the decoder is to exploit already decoded information (e.g., decoded frames) to generate a prediction of the other information still to be decoded. For example, if the frames at display time $n-1$ and $n+1$ visualizing the movement of a person's arm are already decoded, the position of the arm within the (not yet decoded) frame at time index n can be predicted. This prediction is referred to as the *side information*. Due to the complexity of estimating motion, additional error correcting information (such as turbo or LDPC codes) is sent from the encoder to the decoder to allow correcting errors in the side information and complete decoding.

Further author information: (Send correspondence to Jürgen Slowack.)
Jürgen Slowack: E-mail: jurgen.slowack@ugent.be, Telephone: +32 9 331 4957

Applications of Digital Image Processing XXXV, edited by Andrew G. Tescher, Proc. of SPIE
Vol. 8499, 84990L · © 2012 SPIE · CCC code: 0277-786/12/$18 · doi: 10.1117/12.929536

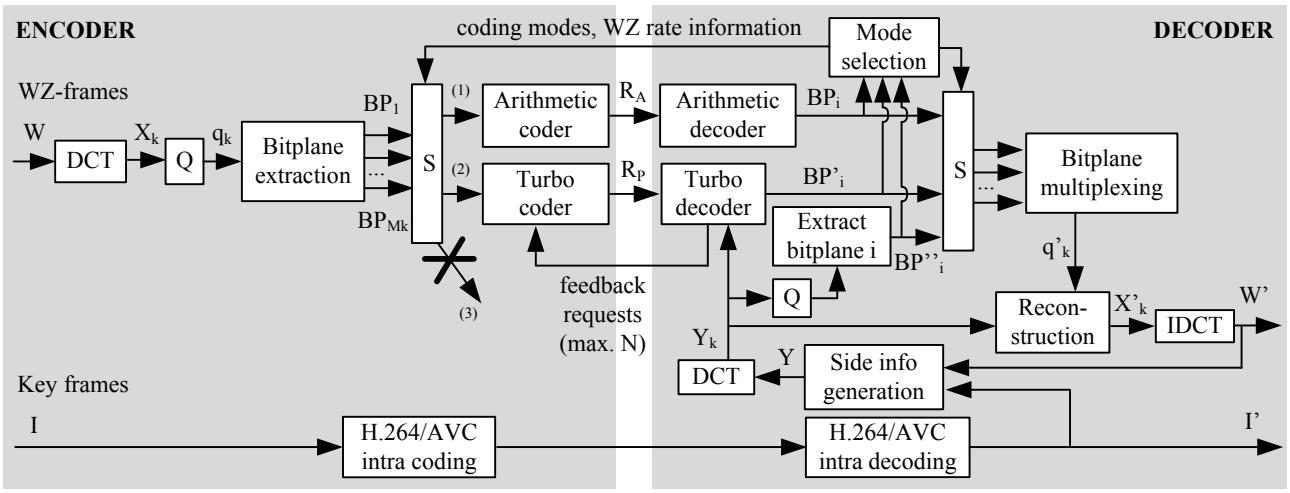

Figure 1. Architecture of the proposed DVC codec.

The amount of error correcting information that needs to be sent from the encoder to the decoder depends on the accuracy of the side information. If the side information is very accurate (as well as its associated correlation model generated for high-efficiency channel decoding[4,5]), few LDPC or turbo codes are required for successful decoding, and vice versa. However, the accuracy of the side information and the correlation noise model are difficult to determine since the original and the side information are not simultaneously available at the encoder or decoder.

Two strategies are typically followed to solve this problem in DVC. In most systems, a feedback channel is used from the decoder to the encoder.[6,7] Through the feedback channel the decoder requests chunks of error correcting information until decoding is considered reliable. The downside of this approach is the additional delay introduced by the feedback channel, as well as the restricted application of such techniques to video streaming scenarios only.

Alternatively, feedback-free DVC systems have been proposed in the literature.[8,9] The encoder in such systems typically generates an estimation of the side information using low-complexity techniques such as averaging of the reference frames or fast motion estimation. However, the downside of this approach is that it adds complexity to the encoder. Also, considering that the decoder-side techniques for side information generation are becoming more and more complex (e.g., due to side information refinement[10,11]), approximating such techniques using low-complexity methods at the encoder will become more and more challenging.

Since in many streaming scenarios a limited form of feedback could be supported, in previous work[12] we proposed a technique to constrain the number of requests through the feedback channel to a fixed number of requests per WZ frame. This number can be defined based on network characteristics, for example. In this paper we extend this approach by exploiting information available from previously decoded WZ frames at other temporal layers.

This paper is structured as follows. In Section 2, details are provided about the DVC codec used as a starting point in this paper. Next, we describe the refined techniques for defining the feedback requests in Section 3. The results are presented in Section 4, followed by conclusions in Section 5.

2. CODEC DESCRIPTION

A schematic diagram of the proposed DVC system is presented in Fig. 1. First, we will describe the operation of the encoder after which details will be provided for the decoder.

At the encoder, the frame sequence is partitioned into key frames and Wyner-Ziv (WZ) frames, using a fixed pattern (e.g., every fourth frame could be a key frame). Key frames are coded without using other frames as

references, i.e., through H.264/AVC intra coding. However, this module is not a specific requirement for the proposed architecture, and could be replaced with any intra codec. The WZ frames, on the other hand, are partitioned into non-overlapping blocks of size 4-by-4 pixels which are coded using the integer approximation of the DCT as used in the H.264/AVC standard. Next, for each WZ frame, coefficients at the same frequency index are grouped together to form coefficient bands. The coefficients in each band are subsequently quantized and bits at corresponding positions are extracted to form bitplanes. For example, all least significant bits of all second DCT coefficients in an entire WZ frame will be grouped to form one bitplane. Finally, depending on the coding mode communicated by the decoder, each bitplane will be either (1) intra coded using a binary arithmetic coder, (2) Wyner-Ziv coded using punctured turbo codes, or (3) discarded (i.e., skip mode).

At the decoder, key frames are decoded without using other frames as references, i.e., using H.264/AVC intra decoding. For each WZ frame, the decoder generates a prediction based on already decoded frames. This prediction (called side information) is generated using the techniques as proposed in DISCOVER.[13] In this technique, the decoder first estimates the motion field between an already decoded past and future frame (in display order). Next, the motion field is refined, smoothed, and interpolated to create the side information (assuming linear motion between the reference frames). In addition to generating the side information, the decoder also estimates the correlation between the side information and the original at the encoder, using techniques from previous work.[14] After generating the side information and modeling the correlation noise, for each bitplane, the decoder estimates the rate required for intra coding (using the binary arithmetic coder), and defines each of the maximum N feedback requests in the case of WZ mode. This step will be discussed in detail in this paper. Using these estimates as well as a rate-distortion model, the decoder decides upon the coding modes to use for each of the bitplanes.[12] Then, the decoder sends the coding modes for all bitplanes to the encoder, in addition to the number of WZ bits required for each of the (maximum) N requests that can be issued in WZ mode. When receiving the requested information from the encoder, depending on the mode, bitplanes are (1) intra decoded through binary arithmetic decoding, (2) WZ decoded following a turbo decoding strategy using the side information and correlation noise model, or (3) skipped. In the latter case, the corresponding side information bitplane is extracted and used as the result. After decoding all bitplanes of a coefficient band, the bitplanes are combined and the coefficients are reconstructed using optimal centroid reconstruction.[15] Finally, after performing the inverse DCT, the decoded WZ frame is retrieved. This frame can be used for generating side information for other frames to be decoded, and so on.

The focus of this paper is on defining each of the maximum N requests through the feedback channel (i.e., for the WZ mode). The goal is to define each of the requests so that the loss in compression efficiency (compared to the case where there is no limit on the number of requests) is as low as possible.

3. PROPOSED METHOD FOR DEFINING THE FEEDBACK REQUESTS

The strategy for defining each of the N WZ requests for each bitplane consists of several steps. First, the decoder estimates the WZ rate required for decoding this bitplane (Sect. 3.1). Next, the error distribution associated with this estimation is approximated (Sect. 3.2). This information is then used to define each of the N requests by jointly minimizing the expected rate overhead of the requests (Sect. 3.3).

After decoding, the minimal number of WZ bit chunks required for decoding is re-estimated (Sect. 3.4). This information is stored so that it can be used in the context of future frames to be decoded.

3.1 Estimating the WZ rate

The (minimal) number of WZ bit chunks required for successful decoding is estimated by partitioning the WZ frames into temporal layers. The WZ frames at each temporal layer have the same distance between the (past and future) reference frames for generating the side information. For example, a group of pictures (GOP) of size four denoted $I_1 - W_2 - W_3 - W_4 - I_5 - W_6 - W_7 - W_8 - I_9 - ...$ with key frames I and WZ frames W consists of two temporal layers. Temporal layer 1 contains the WZ frames $W_2 - W_4 - W_6 - ...$, while temporal layer 2 contains the remaining WZ frames $W_3 - W_7 - ...$, for which decoded key frames are used for generating the side information (e.g., I_1 and I_5 for W_3).

Denote BP_l ($1 \leq l \leq L$) a particular bitplane in the current WZ frame in temporal layer l out of a total of L temporal layers. Denote the collocated bitplanes in the three previously decoded WZ frames as $BP_{l,-1}$, $BP_{l,-2}$, and $BP_{l,-3}$.

To estimate the rate for BP_l, the decoder exploits information available about the previously decoded WZ frames, for which a very accurate post-decoding re-estimation of the WZ rate is available, as will be described in Sect. 3.4. Denote this rate re-estimation as $\hat{R}$. This information is used to estimate the rate for the current bitplane BP_l as

$$R' = \begin{cases} med(\hat{R}_{l,-1}, \hat{R}_{l,-2}, \hat{R}_{l,-3}) & \text{if } l = L, \\ pred(\hat{R}_{l,-1}, \hat{R}_{l,-2}, \hat{R}_{l,-3}) & \text{otherwise,} \end{cases} \tag{1}$$

where *med* denotes the median operator, and *pred* is given by either the minimum, maximum, or median. The latter decision is based on what would have been best for the previously decoded WZ frame at the higher temporal layer. To this end, denote the absolute error for using a particular predictor at the higher temporal layer $l+1$ as:

$$c_{med} = |median(\hat{R}_{l+1,-2}, \hat{R}_{l+1,-3}, \hat{R}_{l+1,-4}) - \hat{R}_{l+1,-1}| \tag{2}$$

$$c_{max} = |maximum(\hat{R}_{l+1,-2}, \hat{R}_{l+1,-3}, \hat{R}_{l+1,-4}) - \hat{R}_{l+1,-1}| \tag{3}$$

$$c_{min} = |minimum(\hat{R}_{l+1,-2}, \hat{R}_{l+1,-3}, \hat{R}_{l+1,-4}) - \hat{R}_{l+1,-1}|, \tag{4}$$

where $\hat{R}_{l+1,-1}$ denotes the post-decoding rate re-estimation for the previously decoded WZ frame at the higher temporal layer $l+1$. For example, in the case of $I_1 - W_2 - W_3 - W_4 - I_5 - W_6 - W_7 - W_8 - I_9 - ...$, the previously decoded frame at the higher temporal layer would be W_7 for W_6 and W_8 (due to hierarchical frame coding).

In essence, the decoder evaluates if the median, maximum, or minimum was optimal for the previously decoded WZ frame at the higher temporal layer, and applies the best predictor as a predictor for the current WZ frame at the lower layer. If $c_{med} \leq c_{max}$ and $c_{med} \leq c_{min}$, then *pred* in Eq. 1 is defined by the median, if $c_{min} \leq c_{med}$ and $c_{min} \leq c_{max}$ the minimum is used, otherwise the maximum is used.

This approach is driven by the fact that a sudden change in motion characteristics will often lead to a sudden change in WZ rate for an entire GOP. Hence, by analyzing which predictor was optimal at the higher temporal layer, the decoder is better able to cope with such motion changes at lower temporal layers.

3.2 WZ rate estimation accuracy

As in our previous work,[12] the difference between the true WZ rate R and its estimation R' as described in the previous subsection, is modeled as a Laplace distribution: $f_{R-R'}(x) = \frac{\alpha}{2} e^{-\alpha|x|}$. The distribution's scale parameter α is estimated during decoding using maximum likelihood fitting of the error samples of the pre-decoding and post-decoding rate estimates R' and $\hat{R}$, respectively. This results in the following equation for the α[12] parameter associated with BP_l:

$$\alpha = \frac{1}{\frac{1}{M} \sum_{m=1}^{M} |\hat{R}_{l,-m} - R'_{l,-m}|}, \tag{5}$$

where M has been defined equal to 10 as in our previous work. To avoid very small or very large values, α is clipped so that it corresponds to a variance in the interval $[0.1; 20]$.

3.3 Defining the N requests

The N requests are defined using the results from WZ rate estimation and error distribution modeling, as detailed above. At the final request, WZ decoding should be successful with very high probability. Otherwise, decoding failures will introduce a high performance penalty due to the overhead in bit rate and the quality degradation of the output frame. This quality degradation may propagate to other frames when side information for these frames is generated using the quality degraded frame as a reference.

In our previous work, to guarantee successful decoding at request N, the total number of bit chunks $\bar{R}^N$ received for BP_l at the final request is defined using a probability threshold ϵ:[12]

$$\bar{R}^N = R' - \frac{\ln 2\epsilon}{\alpha}, \tag{6}$$

where α and R' are defined as described in the previous subsections, and ϵ is taken equal to 0.1%. Here, we refine this approach by exploiting information from other temporal layers. Due to the fact that the distance between the reference frames is larger at a higher temporal layer, the rates are also higher. For a bitplane BP_l in the current WZ frame and its collocated bitplane $BP_{l+1,-1}$, experiments revealed that the following property holds with high probability: $R_l \leq R_{l+1,-1}$. This fact is exploited to refine the final request, resulting in:

$$\bar{R}^N = \min\left(\hat{R}_{l+1,-1}, R' - \frac{\ln 2\epsilon}{\alpha}\right), \tag{7}$$

Having defined the final request, the remaining $N-1$ requests are defined by minimizing the expected rate overhead for all requests:

$$arg \min_{\{\bar{R}^1,\cdots,\bar{R}^{N-1}\}} \sum_{i=1}^{N} \int_{\bar{R}^{i-1}}^{\bar{R}^{i+1}} f_{R-R'}(x-R') \cdot (\bar{R}^i - x)dx, \tag{8}$$

with $\bar{R}^0$ defined zero. This optimization problem is solved during decoding using numerical techniques.

3.4 Post-decoding rate estimation

WZ decoding will typically be successful at a particular request K (with $1 \leq K \leq N$) but unsuccessful at request $K-1$. Using the decoded bitplane, the decoder can adopt a turbo coding-decoding procedure to obtained the minimal number of WZ bit chunks required for decoding within the interval $\left(\bar{R}^{K-1}, \bar{R}^K\right]$. This information can then be used in the context of rate estimation for future WZ frames to be decoded. In rare cases of unsuccessful decoding at request N, $\hat{R}$ is set equal to $\bar{R}^N$ plus five bit chunks.

4. RESULTS

Tests have been conducted on eight different sequences: Foreman, Table Tennis, Mother and Daughter, Bus, Coastguard, Silent, Stefan, and Mobile Calendar. All sequences are in CIF resolution, 30 Hz, coded with a GOP of length four. Only the luma component is coded to enable comparing with DISCOVER.[13]

First, the proposed system is compared to our previous work for defining the feedback requests.[12] Table 1 provides an overview of the average Bjøntegaard delta[16] rate difference compared to our previous work,[12] for different values of N. From these results it is clear that the proposed technique provides significant gains over our previous approach. The main reason is that exploiting inter-layer correlations enables better dealing with sudden motion changes. This decreases the rate overhead of the N requests, and, allows for better mode decision by using more accurate intra and WZ rate estimates.

Fig. 2 and Fig. 3 provide a comparison between the proposed system, the DISCOVER DVC codec, and H.264/AVC intra and inter coding. For the latter, the extended profile was used, RDO enabled, one slice per picture, coded using a fixed quantization parameter. As illustrated by the results, even with feedback constraints the proposed system is able to slightly outperform the unconstrained DISCOVER codec. This is due to improvements from previous work as well as techniques proposed in this paper. Compared to H.264/AVC, the proposed system is able to outperform H.264/AVC intra but falls behind on H.264/AVC inter coding. Results for other sequences lead to similar conclusions.

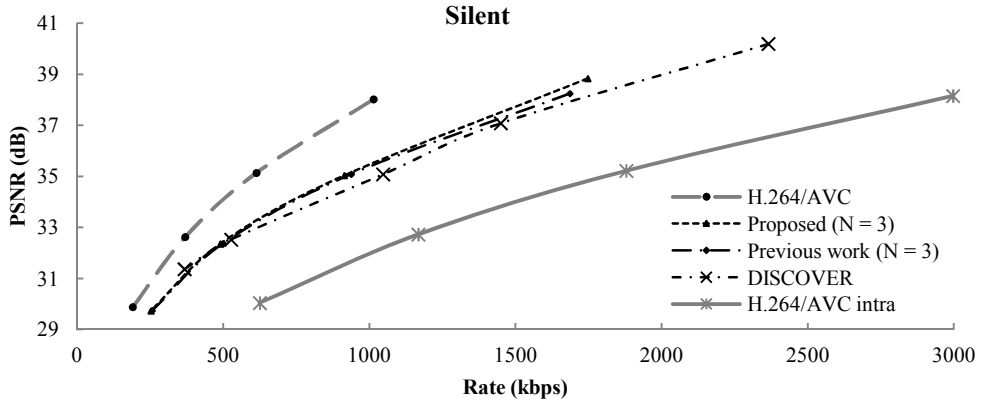

Figure 2. Comparison between our system with $N = 3$, our previous work with $N = 3$,[12] the unconstrained DVC system of DISCOVER,[13] H.264/AVC intra and H.264/AVC inter coding, for the Silent sequence (CIF, GOP 4).

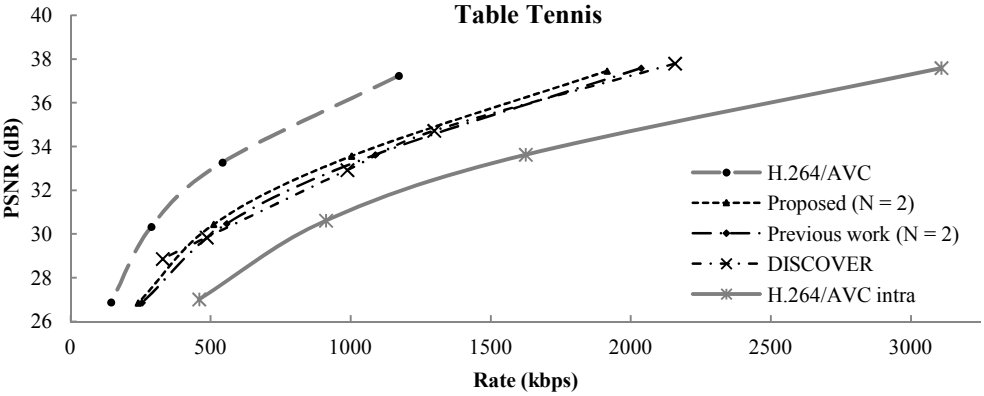

Figure 3. Comparison between our system with $N = 2$, our previous work with $N = 2$,[12] the unconstrained DVC system of DISCOVER,[13] H.264/AVC intra and H.264/AVC inter coding, for the Table Tennis sequence (CIF, GOP 4).

Table 1. Average Bjøntegaard delta[16] rate difference compared to our previous work,[12] for a GOP of size 4 (negative values indicate compression gain).

	$N = 3$	$N = 2$	$N = 1$
Foreman	-1.4 %	-2.5 %	-4.3 %
Table Tennis	-2.3 %	-6.6 %	-11.4 %
Mother Daughter	-1.2 %	-4.5 %	-9.8 %
Bus	-0.2 %	-0.5 %	-1.1 %
Coastguard	-0.2 %	-1.9 %	-4.7 %
Silent	-1.8 %	-3.9 %	-8.2 %
Stefan	-0.8 %	-1.2 %	-1.9 %
Mobile Calendar	0.1 %	-1.7 %	-5.5 %

5. CONCLUSIONS

The use of a feedback channel in DVC has significant implications on system delay. To overcome this issue, feedback-free DVC systems have been proposed in the literature. However, the problem with such systems is that they add complexity to the encoder, which is not desired. Given that a limited form of feedback may be supported in many streaming scenarios, in this paper we proposed a refined method for defining the feedback requests when the total number of such requests is constrained to a fixed maximum. This technique refines our previous work by exploiting correlations between temporal layers, which results in significant gains.

ACKNOWLEDGMENTS

The research activities that have been described in this paper were funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT-Flanders), the Fund for Scientific Research-Flanders (FWO-Flanders), and the European Union.

REFERENCES

[1] F. Pereira, L. Torres, C. Guillemot, T. Ebrahimi, R. Leonardi, and S. Klomp, "Distributed Video Coding: Selecting the most promising application scenarios," *Signal Processing : Image Communication* , pp. 339–352, 2008.

[2] X. Guo, Y. Lu, F. Wu, D. Zhao, and W. Gao, "Wyner-ziv-based multiview video coding," *IEEE Transactions on Circuits and Systems for Video Technology* **18**, pp. 713 –724, june 2008.

[3] S. Rane, P. Baccichet, and B. Girod, "Systematic lossy error protection of video signals," *IEEE Transactions on Circuits and Systems for Video Technology* **18**, pp. 1347 –1360, Oct. 2008.

[4] J. Škorupa, J. Slowack, S. Mys, P. Lambert, and R. Van de Walle, "Accurate correlation modeling for transform-domain Wyner-Ziv video coding," in *Proc. Pacific-Rim Conference on Multimedia (PCM)*, pp. 1–10, December 2008.

[5] C. Brites and F. Pereira, "Correlation noise modeling for efficient pixel and transform domain Wyner-Ziv video coding," *IEEE Trans. Circuits Syst. Video Technol.* **18**, pp. 1177–1190, September 2008.

[6] A. Aaron, S. Rane, E. Setton, and B. Girod, "Transform-domain Wyner-Ziv codec for video," in *Proc. SPIE Visual Communications and Image Processing*, **5308**, pp. 520–528, January 2004.

[7] C. Brites, J. Ascenso, J. Q. Pedro, and F. Pereira, "Evaluating a feedback channel based transform domain wyner-ziv video codec," *Signal Processing: Image Communication* , pp. 269–297, 2008.

[8] M. Morbée, J. Prades-Nebot, A. Pizurica, and W. Philips, "Feedback channel suppression in pixel-domain Distributed Video Coding," in *Annual Workshop on Circuits, Systems and Signal Processing (ProRISC)*, pp. 154–157, November 2006.

[9] C. Brites and F. Pereira, "An efficient encoder rate control solution for transform domain wyner-ziv video coding," *IEEE Transactions on Circuits and Systems for Video Technology* **21**, pp. 1278–1292, September 2011.

[10] 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* **19**, pp. 1327–1341, september 2009.

[11] W. Liu, L. Dong, and W. Zeng, "Motion refinement based progressive side-information estimation for wyner-ziv video coding," *IEEE Transactions on Circuits and Systems for Video Technology* **20**, pp. 1863 –1875, dec. 2010.

[12] J. Slowack, J. Skorupa, N. Deligiannis, P. Lambert, A. Munteanu, and R. Van de Walle, "Distributed video coding with feedback channel constraints," *IEEE Transactions on Circuits and Systems for Video Technology* **22**, pp. 1014 –1026, july 2012.

[13] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret, "The DISCOVER codec: Architecture, techniques and evaluation," in *Proc. Picture Coding Symposium (PCS)*, November 2007.

[14] J. Slowack, S. Mys, J. Škorupa, P. Lambert, C. Grecos, and R. Van de Walle, "Accounting for quantization noise in online correlation noise estimation for distributed video coding," in *Proc. Picture Coding Symposium (PCS)*, May 2009.

[15] D. Kubasov, J. Nayak, and C. Guillemot, "Optimal reconstruction in Wyner-Ziv video coding with multiple side information," in *IEEE MultiMedia Signal Processing Workshop*, pp. 183–186, October 2007.

[16] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," April 2001. VCEG Contribution VCEG-M33.