THE DISCOVER CODEC: ARCHITECTURE, TECHNIQUES AND EVALUATION

X. Artigas[1], J. Ascenso[2], M. Dalai[3], S. Klomp[4], D. Kubasov[5], M. Ouaret[6]

[1]Technical University of Catalonia, [2]Instituto Superior Técnico, [3]University of Brescia,
[4]Leibniz Universität Hannover, [5]INRIA Rennes, [6]Ecole Polytechnique Fédérale de Lausanne

ABSTRACT

Distributed Video Coding is becoming more and more popular among the research community, because of its interesting theoretical contributions and because there are still many open problems waiting to be solved. This paper introduces the codec architecture and the associated tools adopted by DISCOVER (DIStributed COding for Video sERvices), a European project* which has been devoted to the advancement of Distributed Video Coding for two years. Along with the general description and pointers to references with more detailed information, this paper also presents some of the results obtained with the DISCOVER codec. An extended performance analysis and the codec's executable file are both publicly available on the project's web site www.discoverdvc.org.

Index Terms- Wyner-Ziv Coding, Distributed Video Coding

1. INTRODUCTION

Recently, intense research has been conducted in the field of Distributed Video Coding (DVC), a new paradigm for video compression. DVC is the consequence of information-theoretic results obtained by Slepian and Wolf [1] for lossless distributed source coding (DSC), and by Wyner and Ziv [2] for the lossy case. In short, their theorems state that, under same conditions, the rate-distortion performance achieved when performing joint encoding and decoding of two correlated sources can also be obtained by doing separate encoding and joint decoding. In this latter case, it is the task of the decoder to exploit the correlation between the sources to code, meaning that the complexity balance between encoder and decoder can be potentially reversed with respect to traditional coding methods. For instance, in video coding, the very significant burden of motion estimation can be (fully or partially) shifted from the encoder to the decoder. As a result, DVC is recently becoming very appealing for a wide range of real life applications where the computational power, memory and/or battery are scarce at the encoder, such as visual sensor networks, disposable video cameras or multiview image acquisition systems.

First practical implementations of DVC systems were made in [3] and [4]. In [3], the PRISM codec is introduced, which is based on independent syndrome coding of pixel blocks. In [4], a codec based on turbo codes operating on the whole frame is proposed.

In this paper, the DISCOVER monoview codec and the most significant techniques it uses are described along with the codec's performance, evaluated for representative sequences and compared to well-known standard video codecs. Moreover, an implementation of the codec is publicly available on the project's web page [5]. Although the DISCOVER project also developed a DVC multiview video codec, this paper describes only the monoview one.

This paper is structured as follows. The DISCOVER architecture is first presented in Section 2. The encoder-specific tools are then described in Section 3, and the decoder-specific tools in Section 4. The rate-distortion (RD) performance of the proposed codec is evaluated next in Section 5, and finally, some concluding remarks are drawn in Section 6.

2. DISCOVER CODEC ARCHITECTURE

The codec architecture, whose block diagram is depicted in Figure 1, is based on the scheme proposed in [6]. Regarding the scheme described in [6], however, many techniques have been added or improved within the project, for example, to enhance the performance of basic building blocks (e.g., Section 4.1) or to cope with problems associated to on-line estimation of parameters (e.g., Sections 3.4, 4.2, 4.3).

The encoding phase, whose operations are represented by blocks 1 to 3, is very simple. Block 1 is devoted to the splitting of the incoming frame sequence in two parts. A first set of frames, called key frames, is encoded with conventional techniques, namely by an H.264/AVC encoder operating in intra-mode (Block 2). The remaining frames, which are the Wyner-Ziv (WZ) frames, are instead encoded in a distributed fashion by Block 3 as described in the following.

First, every WZ frame undergoes a block based transform (Block 3a), and the obtained transformed coefficients are quantized (Block 3b). These coefficients are then organized in bands where every band contains the coefficients associated to the same frequency in different blocks. The bits representing these coefficients are ordered bit plane by bit plane and are fed into a systematic channel encoder (Block 3c), which computes a set of parity bits representing the syndrome of the encoded bit planes (systematic bits are discarded). These bits are stored in a buffer (Block 3d) and progressively transmitted to the decoder which iteratively asks for more bits during the decoding operation, using the feedback channel. Block 3e computes an initial number of bits to transmit (R_{min}) for each bit plane and band, in order to avoid the high number of iterations (and thus delay and decoding complexity) inevitably required if this minimal number of bits is not always transmitted first.

The decoding process, represented by blocks 6 to 10, is more complex, due to the fact that the temporal correlation is exploited (by motion estimation) and modeled in this phase. Conventionally encoded key frames are first decoded by Block 4, typically an H.264/AVC decoder. These frames are then used in Block 5 for the construction of the so-called side information (SI), which is an estimate of the original WZ frames. In order to produce the SI for a given WZ frame, a motion compensated interpolation between the two closest reference frames (adjacent frames for GOP=2) is performed. The difference between the original WZ frame and the corresponding SI can be considered as correlation noise in a virtual channel; a Laplacian model is used to obtain a good approximation of the residual (WZ – SI) distribution [7].

*The work presented was developed within DISCOVER, a European project (http://www.discoverdvc.org), funded under the European Commission IST FP6 programme.

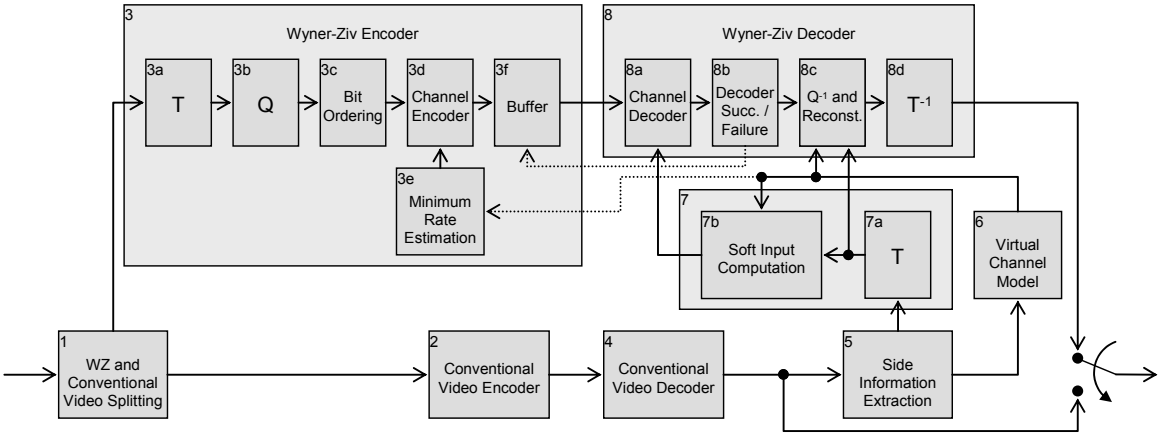

Figure 1. Block diagram of the DISCOVER architecture. Dotted lines represent the feedback channel.

In Block 7, the same transform used at the encoder is applied to the SI and an estimate of the coefficients of the WZ frame is thus obtained. From these coefficients, soft values for the information bits are computed, taking into account the statistical modeling of the virtual noise, which is created in Block 6. These soft values are fed to Block 8, which performs the proper Wyner-Ziv decoding. Here, Block 8a operates the channel decoding, whose success or failure is established by Block 8b using adequate criteria. If the decoding fails, i.e. if the received parity bits are not enough to guarantee successful decoding with a low bit error rate, then more parity bits are requested using the feedback channel. This is iterated until successful decoding is obtained. In this case, Block 8c uses the decoded bits to obtain the reconstructed coefficients using the virtual channel model estimated in Block 6 and the SI coefficients. After Block 8d inverts the transform applied by Block 3a, the decoded video sequence is obtained by conveniently multiplexing the decoded key frames and WZ frames.

3. ENCODER TECHNIQUES

3.1. WZ and Conventional Video Splitting

One common approach in the DVC paradigm is to perform frame interpolation at the decoder to create the SI frames using successive groups of a fixed number of pictures, i.e. using a fixed GOP (Group of Pictures). However, the varying temporal correlation in the sequence can be better exploited according to the content characteristics if varying GOP sizes are used. This implies using longer GOPs when there is more temporal redundancy (the amount of motion is low or motion is better behaved) and shorter GOPs when there is less temporal redundancy (the amount of motion is high or motion is badly behaved). For the DISCOVER codec, an adaptive GOP size selection module was developed (Block 1) to control the (non-periodic) insertion of key frames in between the WZ frames in an adaptive way [8]. This technique has a low complexity and can be divided into two parts: 1) Activity measures: simple but powerful metrics are used to evaluate the activity along the video sequence making use of low level features, notably histogram based; 2) GOP length decision: a new hierarchical clustering algorithm was developed to non-periodically temporally segment the sequence. This algorithm groups frames with similar motion content in order to construct GOPs which are more correlated. The adaptive GOP size allows achieving a better overall rate-distortion performance when compared to a rigid, fixed GOP size approach [8].

3.2. Transform and Quantization

The Wyner-Ziv frames are first transformed using a 4×4 discrete cosine transform (DCT) whose coefficients are organized in 16 bands $b \in [1, 16]$. The first band $b = 1$ containing low frequency information is often called the DC band, to distinguish it from the others which are called the AC bands. Each DCT band b is quantized separately using a predefined number 2^{M_b} of levels, depending on the target quality for the WZ frame.

A uniform scalar quantizer is used for the DC band, assuming the data range $[0, 2^{11})$. This means that the range for the q-th quantization interval is $I_{DC}^{q} = [qW_{DC}, (q+1)W_{DC})$, where $W_{DC} = 2^{(11-M_1)}$ is the DC quantization step size, and M_1 is the number of bits reserved for each quantized value of the first (DC) band.

For AC bands, a dead-zone quantizer with doubled zero interval is applied. The dynamic data range $[-\mathrm{MaxVal}_b, \mathrm{MaxVal}_b)$ is calculated separately for each b-th band, $b > 1$, to be quantized, and transmitted to the decoder inside the encoded bit stream. The quantization step size in this case is $W_b = \left\lceil \frac{2 \cdot \mathrm{MaxVal}_b}{2^{M_b}} \right\rceil$, and the quantization intervals are defined as follows:

$$I_b^q = \begin{cases} [(q-1)W_b, qW_b) & q < 0 \\ [-W_b, W_b) & q = 0 \\ [qW_b, (q+1)W_b) & q > 0 \end{cases} \quad (1)$$

The quantization indices q of each DCT band b are then organized in M_b bit planes and fed to the channel encoder.

3.3. Channel Encoder and Buffer

The channel encoder, also known as the Slepian-Wolf encoder, uses the rate-compatible LDPC Accumulate (LDPCA) codes for distributed source coding introduced in [9]. The LDPCA codes better approach the capacity of a variety of communication channels, including the virtual channel in DVC, when compared to the turbo codes [9]. An LDPCA encoder consists of an LDPC syndrome-former concatenated with an accumulator. For each bit plane, syndrome bits are created using the LDPC code and accumulated modulo 2 to produce the accumulated syndrome. The Wyner-Ziv encoder stores these accumulated syndromes in a buffer and initially transmits only a few of them in chunks. These initial chunks correspond to the minimal theoretical rate $R_{\min}$, which is discussed in the next subsection. If the Wyner-Ziv decoder fails, more accumulated syndromes are requested from the encoder buffer using a feedback channel. To aid the decoder detecting residual errors, an 8-bit CRC sum of the encoded bit plane is also transmitted.

3.4. Minimum rate estimation

In order to reduce the number of accumulated syndrome requests to be made by the decoder (which has a strong impact on decoding complexity), the encoder can estimate a minimum number of accumulated syndromes to be sent per bit plane and per band. The method adopted in the DISCOVER codec is based on the Wyner-Ziv rate-distortion bound [2] for two correlated Gaussian sources. This bound defines the minimal rate at which one source (X) can be transmitted at a given distortion D^X, to be $R_{min}(D^X) = \frac{1}{2}\log_2 \frac{\sigma^2}{D^X}$, where σ^2 is the variance of the correlation noise between the two sources, given that the second source (Y, the SI) is known perfectly at the decoder. A separate rate for each bit plane can be obtained by estimating the diminution of the distortion brought by this bit plane with respect to previously decoded bit planes. σ^2 is a parameter of the virtual channel model discussed below. It is estimated at the decoder side and can be sent back to the encoder via the return channel. More details cannot be given here due to space limitations. However, an alternative method is described in [10] and is based on the conditional entropy $H(X|Y)$ between the data to be encoded (X) and the SI (Y). This conditional entropy in turn is expressed using the crossover probability $p_{cr} \equiv \mathbb{P}(\hat{x}_k(y) \neq x_k)$, where $\hat{x}_k(y)$ is the decoder estimation of the k-th bit of the original signal x using side information value y. Finally, the crossover probability is estimated using the virtual channel model, which is made known to the encoder via the return channel as in the first approach.

4. DECODER TECHNIQUES

4.1. Side Information Extraction

The techniques to generate the side information at the decoder influence significantly the rate-distortion performance of the Wyner-Ziv video codec, in the same way as efficient motion estimation and compensation tools have been establishing the compression advances and performance for block based hybrid video coding. However, since the goal in DVC is to find an estimate of the current WZ frame, a different set of tools is needed. There are two major approaches to create side information in WZ video coding: hash-based motion estimation and motion compensated interpolation (MCI). The latter one was selected for the DISCOVER codec due to the more consistent RD results obtained. In MCI, a motion field closer to the true motion is estimated between backward X_b (past) and forward X_f (future) reference frames; then motion compensation between the two references is performed to obtain the side information.
The following techniques [8][11] are used to obtain high quality side information. First, forward motion estimation from X_b to X_f is performed. A block matching based on a modified MAD (mean absolute difference) criterion is used in order to regularize the motion vector field, which favors motion vectors closer to the origin. Then, bidirectional motion estimation is performed in order to find symmetric motion vectors from the current WZ frame to X_b and X_f. Spatial motion smoothing based on a weighted vector median filter is applied afterwards to the obtained motion field to remove outliers. Finally, motion compensation is performed between X_b and X_f along the obtained motion field, so as to generate the side information. A hierarchical coarse-to-fine approach is used in the bidirectional motion estimation: the first iteration corresponds to a large block size (16×16) and tracks fast motion reliably, while the second iteration achieves higher precision using a smaller block size (8×8). The motion search is performed using the half-pixel precision method described in [12].

4.2. Virtual Channel Model and Soft Input Calculation

The DISCOVER codec uses a Laplacian distribution (as in [6][7]) to model the correlation noise, i.e. the error distribution between corresponding DCT bands of SI and WZ frames. The Laplacian distribution parameter is estimated online at the decoder and takes into consideration the temporal and spatial variability of the correlation noise statistics. This technique avoids a common practice in the literature [6] which is to compute the correlation noise distribution (CND) parameters using a training (offline) stage. This offline process is not realistic because it requires either the encoder to recreate the side information (increasing the encoder's complexity) or to have the original data available at the decoder. The techniques used in the DISCOVER codec (based on [7]) estimate the Laplacian distribution parameter α at the DCT band level (one α per DCT band and frame) and at the coefficient level (one α per DCT coefficient). The estimation approach uses the residual frame, i.e. the difference between X_b and X_f (along the motion vectors), as a confidence measure of the frame interpolation operation, and also a rough estimate of the side information quality. The Laplacian distribution model is then used to convert the side information DCT coefficients into soft-input information to the LDPC decoder. The conditional probability $\mathbb{P}(\mathrm{WZ}|\mathrm{SI})$ obtained for each DCT coefficient is converted into conditional bit probabilities by considering the previously decoded bitplanes and the value of the side information.

4.3. Channel Decoder and Decoder Success/Failure

Given the soft-input information calculated using the virtual channel model, the channel decoder (also known as the Slepian-Wolf decoder) corrects the bit errors in the side information using a belief propagation procedure on the initial number of accumulated syndromes corresponding to R_{min}, received from the encoder buffer. To establish if decoding is successful the convergence is tested by computing the syndrome check error, i.e. the Hamming distance between the received syndrome and the one generated using the decoded bit plane, followed by a cyclic redundancy check (CRC) [10]. If the Hamming distance is different from zero, then the decoder proceeds to the next iteration. After a certain amount of iterations (experimentally it was found that 100 is enough), if the Hamming distance remains different from zero, then the bit plane is assumed to be erroneously decoded and the LDPCA decoder requests for more accumulated syndromes via the return channel. If the Hamming distance is equal to zero, then the successfulness of the decoding operation is verified using the 8-bit CRC sum [10]. If the CRC sum computed on the decoded bit plane matches the value received from the encoder, the decoding is declared successful and the decoded bit plane is sent to the reconstruction module. Otherwise the decoder requests more accumulated syndromes and thus a final low error probability is always guaranteed. For more information about LDPCA decoding refer to [9].

4.4. Reconstruction and Inverse Transform

The decoded value $\hat{x}$ is reconstructed in a mean squared error-optimal way as the expectation of x given the decoded quantization index q and the side information value y: $\hat{x} = \mathbb{E}\{x|q, y\}$. The calculation of this expectation value is performed using closed-form expressions derived in [13] for a Laplacian correlation model. Those frequency bands for which no information was transmitted from the encoder (the number of quantization levels is zero), are taken directly from the SI. After that, the inverse 4x4 DCT transform is applied, and the whole WZ frame is restored in the pixel domain.

5. EXPERIMENTAL RESULTS

The following test conditions have been used to obtain the example rate-distortion (RD) results presented here: All frames of "Hall & Monitor" and "Foreman" sequences, QCIF@15Hz with a GOP length of 2 one of the most common GOP sizes evaluated in the literature [7][10-13]. For the motion interpolation, ±32 pixels are used for the search range of the forward motion estimation; both references are first low pass filtered with a 3×3 size mean filter. Key frames are always encoded with H.264/AVC Intra (Main profile), and the quantization parameters (QP) for each RD point are chosen so that the average quality (PSNR) of the WZ frames is similar to the quality of the key frames. All rate and distortion results refer only to the luminance.

The DISCOVER codec is compared with standard low complexity encoders (although the DISCOVER codec has even lower encoding complexity). On one hand, H.263+ Intra and H.264/AVC Intra, since they are two very well known codecs where no temporal correlation is exploited (but remind that H.264/AVC Intra exploits quite efficiently the spatial correlation at the cost of some complexity). On the other hand, H.264/AVC with no motion (IBI GOP structure), which has a lower encoding complexity compared to the full H.264/AVC Inter codec since it uses the collocated blocks in the previous and/or future reference frames for prediction (or the Intra mode). Figure 2 presents some DISCOVER codec RD performance results. Significant gains can be observed when compared to H.263+ Intra (over 8dB for "Hall & Monitor"); for H.264/AVC Intra, the performance difference is between -0.5 (high bitrates for "Foreman" sequence) to 3dB (low bitrates for "Hall&Monitor"). Therefore, it is possible to conclude that the DISCOVER WZ codec can exploit the temporal correlation in an efficient way while using a rather simple encoder and still be competitive when compared to the (more complex) H.264/AVC Intra encoder. However, when compared to the H.264/AVC with no motion codec, some performance losses are observed which shows there is a need for future improvements to approach the theoretical performance.

More performance evaluation experiments have been carried out which are not detailed here due to space limitations. However, they are described, along with the obtained performance results and the codec in executable format, on the project's web page [5].

6. CONCLUSIONS

This paper has introduced the codec architecture and the most significant techniques adopted by the DISCOVER monoview WZ codec developed in the DISCOVER project. The experimental results show that the presented codec is already RD competitive when compared with other codecs with similar (low) encoding complexity, even though there is still more room for research. Finally, the complete set of experimental conditions and results and the codec in executable format are publicly available [5] to aid the comparison of future technical developments from all the DVC research community.

7. REFERENCES

[1] J. Slepian and J. Wolf, "Noiseless Coding of Correlated Information Sources", IEEE Trans. on Information Theory, vol. 19, no. 4, 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, January 1976.

[3] R. Puri and K. Ramchandran, "PRISM: A New Robust Video Coding Architecture Based on Distributed Compression Principles", Proc. Allerton Conf., October 2002.

[4] A. Aaron, R. Zhang and B. Girod, "Wyner-Ziv Coding of Motion Video", Proc. Asilomar Conference on Signals and Systems, Pacific Grove, CA, Nov. 2002. Invited Paper.

[5] www.discoverdvc.org

[6] Bernd Girod, Anne Aaron, Shantanu Rane and David Rebollo-Monedero, "Distributed Video Coding", Proceedings of the IEEE, vol. 93, no. 1, January 2005.

[7] C. Brites, J. Ascenso, F. Pereira, "Studying Temporal Correlation Noise Modeling for Pixel Based Wyner-Ziv Video Coding", IEEE International Conference on Image Processing, Atlanta, USA, October 2006.

[8] J. Ascenso, C. Brites and F. Pereira "Content Adaptive Wyner-Ziv Video Coding Driven by Motion Activity", IEEE International Conference on Image Processing, Atlanta, USA, October 2006.

[9] D. Varodayan, A. Aaron and B. Girod, "Rate-Adaptive Codes for Distributed Source Coding", EURASIP Signal Processing Journal, Special Section on Distributed Source Coding, vol. 86, no. 11, November 2006.

[10] D. Kubasov, K. Lajnef and C. Guillemot, "A Hybrid Encoder/Decoder Rate Control for Wyner-Ziv Video Coding with a Feedback Channel", Int. Workshop on Multimedia Signal Processing, Crete, Greece, October 2007.

[11] J. Ascenso, C. Brites and F. Pereira, "Improving Frame Interpolation with Spatial Motion Smoothing for Pixel Domain Distributed Video Coding", 5th EURASIP Conference on Speech and Image Processing, Multimedia Communications and Services, Smolenice, Slovak Republic, July 2005.

[12] S. Klomp, Y. Vatis and J. Ostermann, "Side Information Interpolation with Sub-pel Motion Compensation for Wyner-Ziv Decoder", Int. Conf. on Signal Processing and Multimedia Applications, Setúbal, Portugal, August 2006.

[13] D. Kubasov, J. Nayak and C. Guillemot, "Optimal Reconstruction in Wyner-Ziv Video Coding with Multiple Side Information", Int. Workshop on Multimedia Signal Processing, Crete, Greece, October 2007.

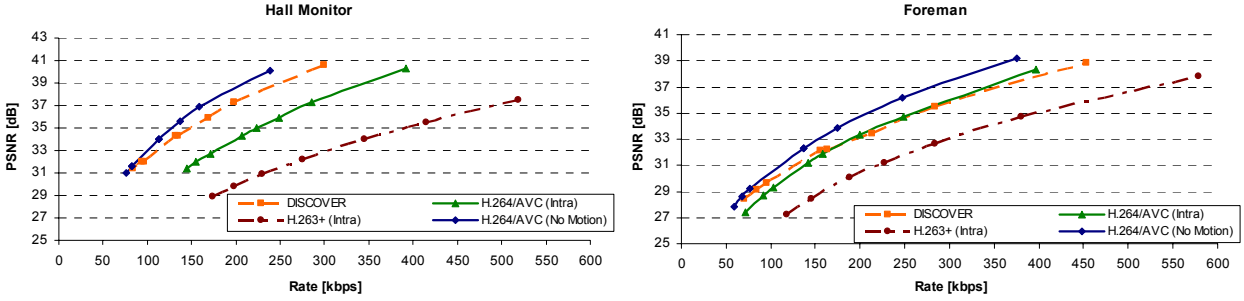

Figure 2. DISCOVER codec Rate-Distortion results.