Distributed Video Coding: Basics, Codecs and Applications

C. Guillemot†, F. Pereira‡, L. Torres*, T. Ebrahimi◇,
R. Leonardi⋆, J. Ostermann△

I. Introduction

It is a recognized fact that a growing percentage of the world population uses nowadays image and video coding technologies on a rather regular basis. These technologies are behind the success and quick deployment of services and products such as digital pictures, digital television, DVDs, and Internet video communications. Today's digital video coding paradigm represented by the ITU-T and MPEG standards mainly relies on a hybrid of block-based transform and interframe predictive coding approaches. In this coding framework, the encoder architecture has the task to exploit both the temporal and spatial redundancies present in the video sequence which may become a rather complex exercise. As a consequence, all standard video encoders have a much higher computational complexity than the decoder (typically 5 to 10 times more complex), mainly due to the temporal correlation exploitation tools, notably the motion estimation process. This type of architecture is well-suited for applications where the video is encoded once and decoded many times, i.e. one-to-many topologies, such as broadcasting or video-on-demand, where

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

† INRIA, Rennes France, Christine.Guillemot@irisa.fr
‡ Instituto Superior Técnico - Instituto de Telecomunicaçoes, fp@lx.it.pt
∗ Technical University of Catalonia, Barcelona, Spain, luis@gps.tsc.upc.edu
◇ Ecole Polytechnique Fédérale de Lausanne, Lausanne Switzerland, Touradj.Ebrahimi@epfl.ch
⋆ Università degli Studi di Brescia, Brescia, Italy, riccardo.leonardi@ing.unibs.it
△Universitȃt Hannover, Hannover, Germany, ostermann@tnt.uni-hannover.de

the cost of the decoder is more critical than the cost of the encoder.

Distributed source coding (DSC) has emerged as an enabling technology for sensor networks. It refers to the compression of correlated signals captured by different sensors which do not communicate between themselves. All the signals captured are compressed independently and transmitted to a central base station which has the capability to decode them jointly. Tutorials on distributed source coding for sensor networks, presenting the underlying theory as well as first practical solutions, have already been published in the IEEE Signal Processing Magazine in 2002 [1] and 2004 [2]. Video compression has been recast into a distributed source coding framework leading to distributed video coding (DVC) systems targeting low coding complexity and error resilience. Correlated samples (pixels or transform coefficients) from different frames are regarded as outputs of different sensors. A comprehensive survey of first DVC solutions can be found in [3]. This paper is a follow-up of these three previous tutorials. While, for sake of completeness, some basics about DSC are reviewed, the paper focuses on the latest developments for distributed video compression (DVC) for both monoview and multiview set-ups. Potential benefits for a range of applications are discussed and most promising application scenarios are identified.

II. DSC: Theoretical Background

DSC finds its foundation in the seminal Slepian-Wolf (SW) [4] and Wyner-Ziv (WZ) [12] theorems. Due to space limitation, only the main concepts are recalled. For more details, the readers are referred to [1], [2], [3].

A. Slepian-Wolf coding

Let X and Y be two binary correlated memoryless sources to be losslessly encoded. If the two coders communicate (see Fig. 1), it is well known from Shannon's theory that the minimum lossless rate for X and Y is given by the joint entropy $H(X,Y)$. Slepian

and Wolf have established in 1973 [4] that this lossless compression rate bound can be approached with a vanishing error probability for long sequences, even if the two sources are coded separately, provided that they are decoded jointly and that their correlation is known to both the encoder and the decoder. The achievable rate region is thus defined by $R_X \geq H(X|Y)$, $R_Y \geq H(Y|X)$ and $R_X + R_Y \geq H(X,Y)$.

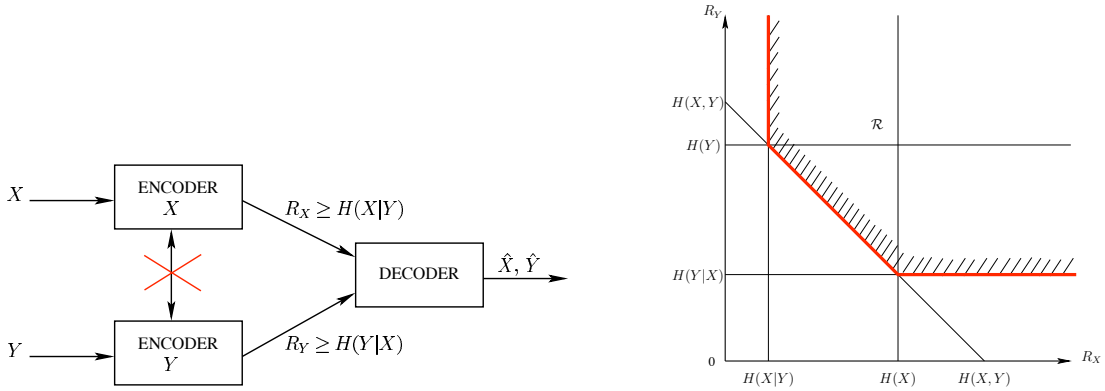

Fig. 1. Distributed coding of statistically dependent i.i.d. discrete random sequences X and Y. Set-up (left); Achievable rate region (right).

The proof of the SW theorem is based on random binning [4], which is non-constructive, i.e., it does not reveal how practical code design should be done. In 1974, Wyner suggested the use of parity check codes to approach the corner points of the SW rate region [5]. These corner points correspond to the asymmetric case where one source is transmitted at full rate (e.g., $R_Y = H(Y)$) and used as side information (SI) to decode the other one (implying $R_X = H(X|Y)$ or reciprocally). The bins partitioning the space of all possible source realizations are constructed from the cosets of the parity check code. The correlation between X and the side information Y is modelled as a "virtual" channel, where Y is regarded as a noisy version of X. Recent channel capacity-achieving codes, block codes [6], turbo codes [7], [8] or LDPC codes [9], have been shown to approach the corner points of the SW region. The compression of X is achieved by transmitting only a bin index, i.e., syndrome or parity bits. The decoder corrects the "virtual" channel noise, and thus

estimates X given the received parity bits and the SI Y regarded as a noisy version of the codeword systematic bits. The achievability of the entire SW rate region, including the symmetric DSC case, has been investigated using linear [10] and LDPC codes [11].

B. Wyner-Ziv coding

In 1976, Wyner and Ziv considered the problem of coding, with respect to a fidelity criterion, of two correlated sources X and Y [12]. They have established the rate-distortion function $R*_{X|Y}(D)$ for the case where the SI Y is perfectly known to the decoder only. For a given target distortion D, $R*_{X|Y}(D)$ in general verifies $R_{X|Y}(D) \leq R*_{X|Y}(D) \leq R_X(D)$, where $R_{X|Y}(D)$ is the rate required to encode X if Y is available to both the encoder and the decoder, and R_X is the minimal rate for encoding X without SI. Wyner and Ziv have shown that, for correlated Gaussian sources and a mean square error distortion measure, there is no rate loss with respect to joint coding and joint decoding of the two sources, i.e., $R*_{X|Y}(D) = R_{X|Y}(D)$. This no rate loss result has been extended in [13] to the case where only the innovation between X and Y needs to be Gaussian, that is where X and Y can follow any arbitrary distribution.

Coding under a fidelity criterion finds its foundations in quantization theory. Practical code constructions based on the Wyner-Ziv theorem thus naturally rely on a quantizer (source code) followed by a SW coder (channel code). WZ coding can thus be regarded as a source-channel coding problem. Under ideal Gaussianity assumptions, the WZ limit can be asymptotically achieved with nested lattice quantizers. Nested lattice constructions are proposed for the WZ coding problem in [14], [15]. The source is first quantized using a fine source code, the coset indexes being then encoded with a SW code which exploits the remaining correlation between the quantized version of X and the SI Y. To minimize the quantization loss (that is to achieve a large granular gain), the lattice quantizer may require high dimensionality, hence high complexity. Solutions based on trellis coded quantizers

[16] are shown to approach the performance of Lattice codes with reduced complexity.

C. From DSC to DVC and potential benefits

The above results suggest that correlated samples (pixels or transform coefficients), taken from different spatial or temporal units, e.g. video frames, and under Gaussianity assumptions, can be quantized and coded independently with minimum loss in terms of rate-distortion (RD) performance with respect to predictive coding, if they are decoded jointly. In predictive coding, motion-compensated frames are used as predictors at the encoder and signalled by transmitting the motion vectors to the decoder. These predictors can be regarded as SI used at both the encoder and the decoder. In DVC, the SI should be constructed and used by the decoder only. Ideally, only the statistical dependence between the WZ encoded samples and the SI needs to be known to the encoder. DVC architectures can therefore potentially present the following attractive features:

- Part of the complexity may be shifted from the encoder to the decoder leading to a flexible trade-off between coder and decoder complexity. This feature may be desirable for a class of applications requiring low encoder complexity or low battery consumption.
- A reduction of the error propagation problem inherent to predictive coding may be expected from the fact that the correlated data (e.g. consecutive frames) will be coded separately (and not predictively).
- In current scalable codecs, upper layers are predicted from lower layers. The refinement signals depend on the encoding used in the lower layers. WZ encoding of samples from upper layers will make the corresponding refinement signals independent of the coding structure or the spatial resolution of the lower layers. This independence between layers also contributes to improved error resilience.
- The DSC principles apply quite naturally to the compression of video sequences captured of the same scene by several cameras. With respect to classical multiview coding

techniques, DVC allows the exploitation of correlation between views without - or with limited - inter-sensor (that is inter-camera) communication.
The above functional benefits are used in Section 4 to identify the most promising DVC applications.

III. DVC: Towards practical solutions

The application of the WZ principles to video compression is not straightforward and requires solving a number of issues. The first issue is the identification in the video sequence of the data to be WZ encoded and the construction of the corresponding SI. The source correlation, one key factor in the RD performance of the systems, greatly depends on the quality of the SI. The theoretic results assume the source correlation to be known at both the encoder and the decoder. However, video signals are highly non-ergodic and thus the source correlation, which is time-varying, needs to be estimated. The accuracy of this estimation will have a strong impact on the compression efficiency. Finally, the WZ limits are shown to be achieved by capacity-approaching codes. Nevertheless, these codes may require long block lengths, which may not be practical for delay-constrained video applications.

A. First DVC architectures

First DVC architectures appeared in 2002 [17], [18] followed by variants, e.g., transposing WZ coding from the pixel to transform domain to better exploit the spatial redundancy. A comprehensive overview of the DVC state-of-the-art in 2004 can be found in [3]. The objectives of low encoding complexity and error resilience are central to the first DVC solutions. The main features of DVC architectures are outlined in the general block diagram shown in Fig.2. Blocks which are grey-shaded as well as the dotted lines correspond to features specific to the two first architectures as explained below.

In PRISM [17], the encoder, based on frame differences, classifies each 16×16 block

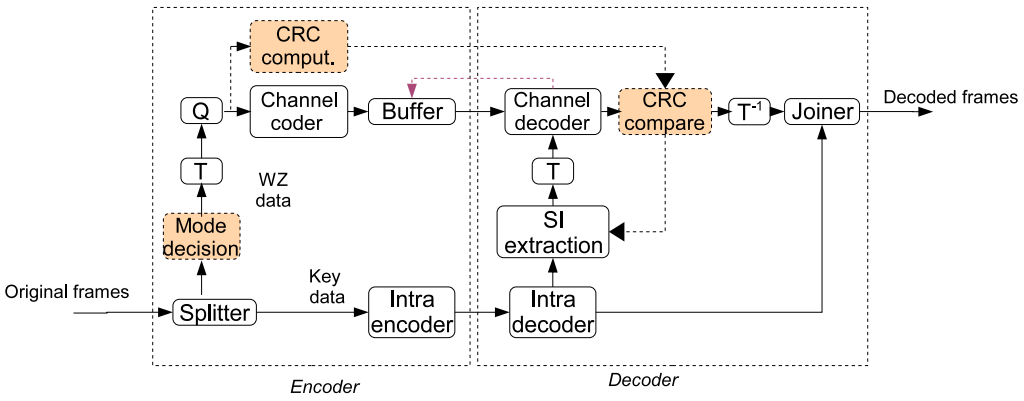

Fig. 2. Distributed video coding architecture.

of the frame into not coded, Intra coded, or WZ coded with a set of fixed rates. The rate chosen for a given block depends on the variance of the frame difference assumed to follow a Laplacian distribution. Each block is transformed using a Discrete Cosine Transform (DCT). Since only the low frequency coefficients have significant correlation with the corresponding estimated block (SI), the high frequency coefficients are Intra coded. The WZ data (low frequency coefficients) are quantised and encoded with a trellis code. Furthermore, the encoder sends a cyclic redundancy check (CRC) word computed on the quantised low frequency coefficients of a block to help motion estimation/compensation at the decoder. A set of motion-compensated candidate SI blocks extracted from previously decoded frames is considered at the decoder. The CRC of each decoded block is compared with the transmitted CRC. In case of deviation, the decoder chooses another motion vector and thus another candidate block. The PRISM codec is shown to be significantly less affected by frame losses than predictive video codecs such as H.263+ [17].

In [18], the WZ coding decision is taken at a frame level, which leads to structuring the sequence into groups of pictures, in which key frames are Intra-coded (typically using a standard codec such as JPEG-2000 or H.264/AVC Intra) and intermediate frames are WZ coded. Each WZ frame is encoded independently of the other frames. The WZ data are

quantised and fed into a punctured turbo coder. The systematic bits are discarded and only the parity bits of the turbo coder are stored in a buffer. The encoder sends only a subset of the parity bits. The SI is constructed via motion-compensated interpolation (or extrapolation) of previously decoded key frames. If the decoder cannot properly decode the current frame, more bits are requested to the encoder via a feedback channel. This allows controlling the bit rate in a more accurate manner and handling changing statistics between the SI and the original frame, at the expense of latency, bandwidth usage, and decoder complexity. The decoder generates the SI (e.g., even frames for a group of pictures of size 2) via motion-compensated interpolation of key frames (e.g. odd frames). After turbo decoding, MMSE estimates of the quantized values, given the received quantization index and the SI, are computed.

B. Recent Technical Developments: Monoview Set-up

The rate-distortion (RD) performance of first DVC solutions was superior to that of H.263+ intraframe coding only (a gain of around 2 dB has been reported in the literature for sequences having low motion such as *Salesman* and *Hall monitor* [3]). However, the significant gap relative to H.263+ motion-compensated interframe coding, and of course that of H.264/AVC, has motivated a lot of recent research efforts. Performance and benefits of WZ coding for layered video compression have also been investigated for scalable compression.

B.1 RD performance improvement

Side information quality: The RD performance of DVC greatly depends on the quality of the SI. In current approaches, the SI, which can be seen as a predictor used at the decoder only, is generated via motion compensated interpolation or extrapolation, often using block-based translational motion models. The decoder computes motion fields between frames, which may be distant from one another. An interpolated version of these motion

fields, typically assuming linear motion, is then used to generate the SI for each WZ frame. This may create misalignments between the SI and the actual WZ frames with a negative impact on the compression efficiency. In addition, motion vector fields obtained this way may suffer from low spatial coherence and may not allow handling covered/uncovered regions.

To cope with these limitations, extra processing is required either at the encoder, at the decoder or on both sides (showing as mentioned previously the DVC flexibility in terms of encoder-decoder complexity budget). As initially suggested in [17], the motion-based extrapolation (or respectively interpolation) steps can be embedded in a multiple motion hypothesis framework. The actual motion vectors are chosen by testing the decoded frames against CRCs or improved hash codes such as a coarse description of blocks within the frame [19], or a low resolution of the video encoded with a zero-motion H.264/AVC codec [20]. This allows improving the motion-based prediction, at the cost of extra bit rate for transmitting the hash codes, as well as of an increased coder and decoder complexity. A choice between forward and backward SI prediction, instead of SI interpolation, is also shown in [21] to better handle covered/uncovered regions. Note that block-based coding mode decision (Intra, WZ coded) at the encoder, already proposed in [17] and further investigated in [22] is another way of going around motion model limitations for handling covered/uncovered regions. This approach comes however with an increased encoder complexity.

The interpolated motion fields can also be refined as parity bits are being decoded, or improved by considering more elaborate motion models. For video sequences of static scenes captured by a moving camera, motion models belonging to the structure-from-motion paradigm coupled with feature point tracking are shown to significantly improve the SI quality [23]. Smoothing techniques removing motion discontinuities at the boundaries and outliers in homogeneous regions are also shown to improve the SI quality [24]. In [25],

the SI is estimated from previously reconstructed video data by formulating the problem as a denoising problem, and then using tools from classical statistical prediction or Wiener theory, avoiding the motion search. This approach, however, requires having a reliable statistical model for the source.

Correlation estimation: The *no loss* result of the Wyner-Ziv theorem comes under the assumption that the statistical dependence (or correlation) between the WZ data and the SI is perfectly known to both the encoder and the decoder, and that it follows a Gaussian distribution. In practice, this statistical dependence needs to be estimated and follows a Laplacian distribution. In first DVC implementations, the parameters of these distributions were computed off-line for each sequence. In [26], a non-stationary model of the noise combining Laplacian and other distributions is considered. The parameters of the Laplacian distribution can be derived from a measure of confidence computed on the motion compensated difference between the key frames. The correlation parameters can then be sent back to the encoder via a return channel, however at the expense of latency and bandwidth usage. Methods to estimate the correlation parameters at the encoder have been described in [27]. An alternate method, similarly to [17], requires the encoder to estimate the SI that will be available at the decoder, thus increasing the encoder complexity. The second approach relies on a model for which the encoder estimates the parameters. It appears clearly that there is a trade-off between encoder complexity, amount of information exchanged between decoder and encoder (via a feedback channel) and accuracy of the correlation parameters used by the encoder.

Rate allocation: Another critical issue in all video compression systems is the rate allocation to the different frames ir blocks for a targeted quality. Rate-distortion models are often used for controlling the quantization parameters. For the WZ coded-data, so far, the only model available is the lower rate bound expressed in terms of the correlation between the two sources and the targeted distortion. However, this lower bound can

be approached only under Gaussian assumptions and if the correlation parameters are perfectly known to both the encoder and decoder. In practice, this value is an underestimation of the actual rate needed to recover the original data for the target distortion. In [18], when the bit error rate at the output of the SW decoder exceeds a given threshold, extra parity bits are requested via a feedback channel. This amounts to controlling the rate of the code by selecting different puncturing patterns at the output of the turbo code. In first implementations, the error rate was assumed to be the true error rate given by the Hamming distance between the (original) sequence to be encoded and the reconstructed sequence, which was not realistic. However, this exact error rate can be replaced, with negligible compression performance loss, by a value estimated from the log likelihood ratio available at the output of the channel decoder. Note that controlling the code rate via puncturing mechanisms is feasible for particular codes (e.g. the turbo codes), but not for others. Punctured LDPC codes perform poorly because the graph resulting from the puncturing contains unconnected and single-connected nodes. LDPC-based rate-adaptive codes are described in [28] and allow for a more flexible rate adaptation with good performance at high compression ratios.

When a feedback channel is not available, rate allocation must be performed at the encoder. One possible solution is to choose a rate among a fixed set of possible rates depending on the correlation with the estimated side information [17], at the expense however of a sub-optimal rate allocation. Here again, there is a trade-off between encoder complexity, information exchanged with possible latency, and accuracy of the rate allocation.

Status in terms of RD performance: The compression performance of current DVC solutions greatly depends on the motion characteristics of the video sequence, hence also of the frame rate of the input sequences. Fast motion will impact negatively the quality of the SI. The quality of the SI also depends on the rate allocated to the Intra-coded data from which it is extrapolated or interpolated. For sequences with low motion and

high frame rates (e.g., 30 Hz), a compression performance gain is achieved with respect to H.264/AVC Intra. However, in presence of fast motion or with low frame rates, this is not always the case.

B.2 Scalable WZ coding

Scalable video coding is attractive for a number of applications. Classical scalable coding solutions are based on closed-loop prediction techniques. Lower layers serve as predictors to the original signal. The refinement signal, a residue of prediction, is then computed. Layered video coding can also be recast into a WZ coding problem, where the upper layers are coded independently and the lower layers are used as side information at the decoder only. Layered distributed coding offers interesting features with respect to closed-loop prediction techniques. The refinement signals do not depend on the coding structure neither from the spatial resolution of the lower layers, leading to reduced encoding complexity and improved error resilience. In contrast with classical closed loop prediction often used in layered representations, there is no need to generate a prediction signal, only the correlation structure needs to be known. The problem of error propagation inherent to predictive approaches is thus alleviated.

Theoretic conditions so that successive refinements of information in a WZ setting can asymptotically achieve the WZ RD function in each layer have been formulated in [29]. Optimum successive refinement can be achieved if the SI Y_{BL} used to decode the base layer and the SI Y_{EL} used to decode the enhancement layer are "equivalent" for a distortion level D_{BL} (see Fig. 3). Equivalence means that, if there exists an auxiliary variable U minimizing the mutual information $I(X;U|Y_{BL})$ for a distortion D_{BL}, then $I(U;Y_{EL}|Y_{BL}) = 0$, i.e., when the side information Y_{BL} for the base layer is given, the side information for the enhancement layer Y_{EL} does not bring additional information on U. Note however that, in practical layered coding systems, this condition is rarely verified.

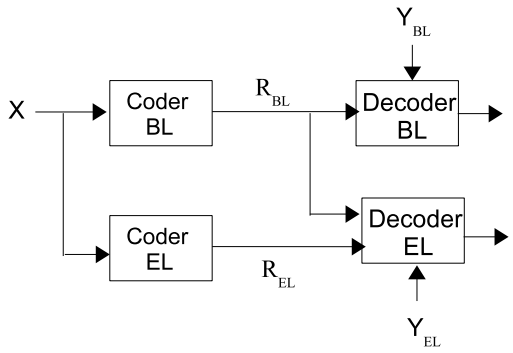

Fig. 3. Successive refinement at the decoder.

A scalable scheme is proposed in [30] where the base layer uses a standard codec and bit planes of the enhancement layers are encoded with a WZ coder formed by a nested quantization followed by LDPC codes. In [30], only the correlation between layers is exploited; the temporal redundancy in the enhancement layers is not exploited. A spatial and temporal scalable codec using PRISM to code the spatial and temporal enhancement layers is described in [31]. Motion vectors from the base layer are used to construct the SI used to decode the enhancement layers, thus exploiting as well temporal correlation. In [27], classical closed-loop inter-layer prediction is combined with WZ coding to exploit the temporal correlation between bitplanes of successive frames. WZ coding is applied on the residues of classical inter-layer prediction.

B.3 WZ coding for error-resilient video transmission

As seen in Section II, WZ coding solutions rely on a quantizer followed by a channel code which aims at correcting errors in the channel modelling the correlation between the WZ encoded data (X) and the SI (Y). The rate of the channel code is chosen so that the code allows the reconstruction of X up to the distortion introduced by the quantizer. In presence of transmission errors, the quality of the SI will degrade. This problem, which can be regarded as a problem of *predictive mismatch* at the decoder side, can be mitigated

by decreasing the rate of the channel code. Parity bits produced by the WZ coder to compress the source X are thus also exploited to combat the mismatch resulting from losses and/or transmission errors [32].

Alternatively, WZ coding can be used as a forward error correction (FEC) mechanism. This idea has been initially suggested in [33] for analog transmission enhanced with WZ encoded digital information. The analog version serves as SI to decode the output of the digital channel. This principle has been applied in [35], [34] to the problem of robust digital video transmission. The video sequence is first conventionally encoded, e.g., using an MPEG coder. The resulting bitstream constitutes the systematic part of the transmitted information which could be protected with classical FEC, possibly with unequal error protection. However, errors in parts of the bitstream, e.g. the temporal prediction residue in conventional predictive coding which may be less heavily protected, may still lead to predictive mismatch and error propagation, if the error correction capability of the code is exceeded. WZ encoded data can be transmitted to facilitate recovery from this predictive mismatch. The WZ data can be seen as extra coarser descriptions of the video sequence, which are redundant if there is no transmission error. The conventionally encoded stream is decoded and the corrupted data is reconstructed using error concealment techniques. The reconstructed signal is then used to generate the SI to decode the WZ encoded data.

Beyond this general formulation, natural questions arise such as the selection of the data to be WZ encoded for achieving the best rate-distortion performance for different channel characteristics. In [35], the residue of the temporal prediction performed by the conventional coder is re-encoded for some frames called *peg* frames. The error propagation is thus confined between two *peg* frames. Such techniques, called systematic lossy error protection techniques, by avoiding the *cliff* effect of conventional FEC, may be advantageous for less critical data to achieve a more graceful degradation of the reconstructed video quality.

C. Recent Technical Developments: Multiview Set-Up

Multiview DVC has only very recently received attention from the scientific community, considering first arrays of cameras capturing static scenes, or light fields. The focus in this paper is on DVC for arrays of video cameras capturing scenes with moving objects. All the techniques discussed above for monoview DVC systems are also needed for multiview DVC. This section focuses on issues specifically involved in multiview DVC mainly related to SI extraction.

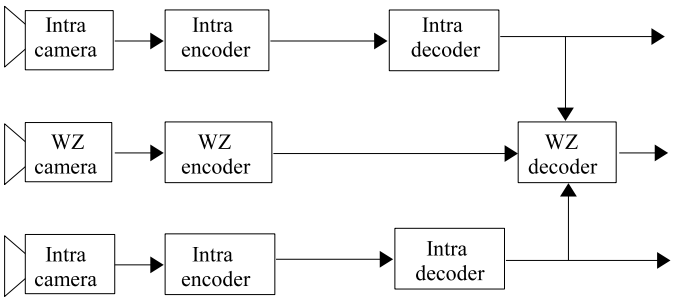

Fig. 4. Multiview DVC illustrative configuration. "Intra-cameras" operate in a conventional fashion while the "WZ-camera" requires joint decoding.

C.1 Capturing inter-camera spatial correlation

To explain the problem of SI extraction in multiview DVC, a simple configuration with three cameras which do not communicate with each other, is considered (Fig. 4). This system can be easily extended to more cameras without fundamentally altering the DVC approach in multiview environments. Two of the cameras are referred to as Intra-cameras and could perform any conventional monoview coding, i.e., their video streams are encoded and decoded independently of the other cameras. The third camera, referred to as WZ-camera, independently encodes the video sequence it captures, but video streams from the other cameras are required for joint decoding (Fig. 4).

As in all multiview compression systems, a major issue is how to predict the signal

captured by the WZ camera from those captured by the neighbouring cameras in an efficient way. While in traditional multiview compression systems, these predictors are computed and used by the encoder and signalled to the decoder, in DVC, they will be computed and used by the decoder only. Hence, the study of solutions based on multiview prediction techniques, but shifted to the decoder side, follows naturally.

A first approach to produce the SI is through Disparity-Compensated View Prediction (DCVP) [36]. DCVP performs frame interpolation based on conventional block-based motion estimation and compensation, but instead of using (temporal) previous and future frames from the same camera, as a monoview decoder would do, it uses reference frames from neighbouring cameras at the same time instances. The motion vectors generated are called Disparity Vectors. It is shown in [36] that the use of disparity-compensated SI in the decoding reduces the bit rate by up to 10% over decoding without SI. In the event of fast motion or low temporal sampling rate, this technique could outperform motion-compensated temporal interpolation which would exploit inner-camera temporal correlation instead of inter-camera spatial correlation. Note that techniques to compensate variations of illumination in the views captured by the different cameras, which are currently used in conventional multiview coding algorithms, could also be naturally shifted to the decoder side to help estimate the SI in multiview DVC.

However, in general, block-based translational models are not sufficiently accurate to predict well spatially adjacent frames because the disparity between objects in different views depends on the distance of the object to the cameras, the cameras positions and the scene geometry. Disparities between views captured by different cameras can also be represented by global models instead of simple block-based translational models. A six-parameter affine model is used in [37], which has been extended to an 8-parameter homography in [38]. The homography is a 3×3 matrix that relates one view to another in the homogenous coordinates system. The translational and affine models are special cases

of the 8-parameter homography model. These models are suitable when the scene can be approximated by a planar surface, or when the scene is static and the camera motion is a pure rotation around its optical center.

The limitations of block-based translational models can also be addressed by introducing an epipolar constraint, which accounts for camera and scene geometry constraints. The epipolar constraint is actually used to reduce the search of correspondences to a 1D problem: given a point in one view, its corresponding point in the other view lies on the epipolar line. Such an approach is considered in [39] in the particular case of a stereo camera set-up in which inner-camera motion vectors are exchanged between sensors. Another approach exploiting camera and scene geometry called View Synthesis Prediction (VSP) has been shown to outperform DCVP by up to 2dB in multiview compression [40]. VSP transposed to the decoder side can address the above limitations of disparity-based view interpolation in multiview DVC. The approach requires first to estimate a depth map to construct the 3D scene on the basis of the neighbouring views, and after to find the parameters and projection matrices associated with the intermediate camera. The 3D scene can then be projected onto the intermediate camera image planes.

C.2 Fusion of temporal and spatial side information

Depending on the position of the cameras, on the spatial and temporal resolutions of the sequences captured, and on the behaviour of moving objects in the scene, temporal correlation in the sequences captured may be higher than inter-camera spatial correlation or vice-versa. The two types of techniques can be combined with advantage. In the sequence captured by the so-called WZ camera in Fig. 4, some data (selected on a block or frame level) will then be coded in Intra mode (e.g. H.264/AVC) while other data use a WZ coder, as explained in section III-B for monoview systems. Fusion techniques may be used to select the adequate temporal or spatial predictions for the WZ coded data, and to

blend them to finally produce the best possible SI. In [37], the 6-parameter warping based prediction is combined with motion-compensated based temporal interpolation while in [38], a fusion is done between temporal side information and homography-based side information. The decision mask is estimated from the best prediction on the previous or following non-WZ frame. Results show PSNR improvements between 0.2 and 0.5 dB when compared to schemes exploiting no fusion, and making use of solely temporal or homographic predictions. However, the gains brought by the homography, or more generally by exploiting inter-camera spatial correlation, depend on the distance between the cameras, on the motion in the scene and on the targeted rate.

Motion vectors estimated between a previous and a current frame captured by an intra-camera can also help estimating motion vectors required to generate the temporal SI used by the WZ decoder (see Fig. 5). Special care needs to be taken to properly transform the motion field from the intra-camera to an appropriate motion field to be used on a different view, for another camera. This is achieved by a matrix transformation. If the cameras lie in the same plane and point in the same direction, this transformation is not needed. This approach called Multi-View Motion Estimation (MVME) is applied in [39] for a stereo set-up.

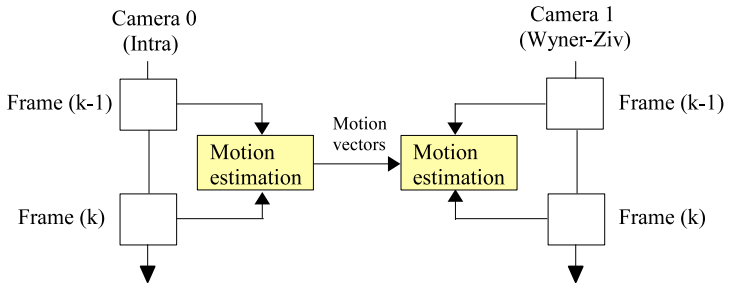

Fig. 5. MVME conceptual scheme: motion vectors are found in the intra camera and used in the WZ camera.

IV. Most promising DVC application scenarios

This section performs an analysis of relevant application scenarios, targeting the identification of the most promising ones from a DVC perspective. The list of scenarios selected for more detailed analysis in this paper is:

1. Wireless video cameras - This application regards the wireless communication/streaming of video signals between remote sites using small, portable cameras for video gathering in diverse situations, e.g. meetings, parties, etc. This type of cameras can be used in embedded systems integrated into cars, trains, airplanes or any mobile environment.

2. Wireless low-power surveillance - Surveillance is the process of monitoring the behaviour of people, objects or processes within systems for conformity to norms in trusted systems, for security or social control using wireless low-power devices. Relevant examples in this area are traffic monitoring, surveillance inside vehicles, home monitoring, wildlife and fire monitoring, and military reconnaissance and monitoring.

3. Video conferencing with mobile devices - Videoconferencing in its most basic form is the transmission of synchronized video and speech back and forth between two or more physically separate locations.

4. Mobile video mail - In contrast to SMS, MMS may have an arbitrary number of attachments of different types, e.g. short text messages, complex documents, pictures or even videos; one possible application of MMS is video mail.

5. Disposable video cameras - Disposable (one-time-use) photo cameras have been around for years and have carved out a healthy niche in the overall photography market. However, nobody had come up with a disposable video camcorder until around June 2005. The business model for this type of camera revolves around the fact that the device will be used by multiple customers, allowing spreading the cost of the hardware over a number of purchases.

6. Visual sensor networks - With the proliferation of inexpensive cameras and non-optical

sensing devices, and the deployment of high-speed, wired/wireless networks, it has become economically and technically feasible to employ a large number of sensing devices for various applications, including on embedded devices. Camera sensor products range from expensive pan-tilt-zoom cameras to high-resolution digital cameras, and from inexpensive web cams and cell phones cameras to even cheaper, tiny cameras.

7. Networked camcorders - Network cameras are devices with acquisition, recording and transmission capabilities (also known as 'camcorders'). For many camcorders, the transmission capabilities can be quite demanding, e.g. transforming the camcorder in a kind of video server. Another possibility is to allow users to remotely control the camcorder in terms of shooting direction/angle, zooming, etc. Network camcorders are shrinking in size and in price, making them feasible for example for people interested in remote monitoring through a local network or the Internet.

8. Distributed video streaming - The huge development of the Internet has given the possibility to realize video streaming systems that allow a user to view a video sequence at its own place while downloading it from a remote server or disk. Following the same idea behind peer to peer networks, it is possible to consider the possibility of performing "distributed streaming"in order to give to the receiver the maximum possible data flow. The video stream is sent to the receiver by different senders in a distributed fashion, reducing the bitrate at the sender sides while increasing it at the receiver also with resilience and reliability advantages.

9. Multiview video acquisition - Most video processing and coding solutions rely on one single camera, referred to as a monoview approach. In the last two decades, extensions to two-camera solutions (stereo) have been investigated with limited success in both coding and video analysis applications. Multiview images of a scene can be used for several applications ranging from free viewpoint television (FTV) to surveillance. In FTV, the user can freely control the viewpoint position of any dynamic real-world scene.

10. Wireless capsule endoscopy - Many human diseases can only be spotted with images of the ill region. With X-ray, the whole body can be photographed but these images are not very accurate, and not all diseases can be detected by this technique. An example is to determine the source of gastrointestinal bleeding since X-ray studies may be unable to pinpoint the exact locations of abnormalities. For this purpose, the capsule endoscopy with image and video capabilities has emerged as an effective way to evaluate gastrointestinal problems.

Table 1 presents a summary of the DVC related potential benefits and drawbacks for the above applications. Table 1 shows that applications such as distributed video streaming and networked camcorders (bidirectional, monoview scenario), wireless low-power surveillance (bidirectional, multiview scenario), wireless video cameras (unidirectional, monoview scenario) and visual sensor networks (unidirectional, multiview scenario) are particularly interesting considering the number of potential DVC benefits.

V. Concluding Remarks

Despite the growing number of research contributions in the past years, there are many open problems to be solved to bring DVC to a level of maturity closer to traditional predictive solutions. In terms of performance, DVC solutions will always be more or less significantly limited by the impossibility to meet the Gaussianity innovation condition set by the WZ theorem. The first major issue to be addressed, both in the mono and multiview set-ups, also follows from the WZ theorem when it assumes that the statistical dependence between the WZ frames and the SI needs to be known by the encoder and decoder for the optimal performance to be reached. Considering that the encoder does not know the SI and the decoder does not know the originals, the estimation by both the encoder and decoder of this statistical correlation will still be a major research item in the years to come since it significantly impacts the RD performance. This problem assumes more nuances

for the multiview set-up, considering the additional number of dimensions involved in the side information creation process and thus on the determination of the noise correlation model(s). Another major DVC challenge is the side information creation process itself due to its strong impact on the final RD performance, notably for less well behaved content and for multiview content. In this context, a major issue to address will be the advantage or not to have the encoder adaptively sending to the decoder auxiliary information and, in the positive case, the type, and amount of this helping data also in relation to its additional encoder complexity. For the multiview case, this issue should be more complex as the relative position of the cameras and the geometry of the scene may have a great impact on the way to efficiently estimate the various side informations. Another important issue regards the study of new channel codes in the Slepian-Wolf module beside the channel codes used until now, mainly LDPCs and turbo codes. Advances in the channel coding arena will very likely also imply advances in the DVC arena. Finally, rate control (notably if no feedback-based architecture is used) is also a critical topic since the way the available rate is invested on the various components of the WZ stream strongly affects the final RD performance.

At some stage, it will be critical to finally understand what the effective functional target of DVC technology is. While it is already theoretically known that RD performance by itself will never be a DVC argument, flexible complexity budgeting, error resilience, scalability and multiview set-ups are the functional candidates for which there are reasonable hopes on specific DVC advantages. Certainly, a functional feature that DVC already shows is the lack of need for communication between the cameras in a multiview set-up but this will only become a functional advantage if DVC RD performance will evolve to outperform standard, low complexity independent camera coding solutions such as those based on H.264/AVC Intra and H.264/AVC Inter with no motion estimation.

Acknowledgement

The authors would like to thank P.Correia (IST), E. Acosta (UPC), X. Artigas (UPC), M. Ouaret (EPFL), F. Dufaux (EPFL), M. Dalai (UNIBS), and S. Klomp (UH) for their contributions to this paper.

References

[1] S.S. Pradhan, J. Kusuma and K. Ramchandran, "Distributed compression in a dense micro-sensor network", *IEEE Signal Processing Magazine*, vol. 19, pp. 51-60, Mar. 2002.

[2] Z. Xiong, A.D. Liveris and S. Cheng, "Distributed source coding for sensor networks", *IEEE Signal Processing Magazine*, pp. 80-94, Sept. 2004.

[3] B. Girod, A. Aaron, S. Rane, D. Rebollo-Monedero, "Distributed video coding", *Proc. IEEE, Special issue on advances in video coding and delivery*, vol. 93, no. 1, pp. 71-83, Invited paper, Jan. 2005.

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

[5] A. Wyner, "Recent results in the Shannon theory", *IEEE Trans. on Information Theory*, vol. 20, No. 1, pp. 2 - 10, Jan. 1974.

[6] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): Design and construction", *Proc. IEEE Data Compression Conference*, Snowbird, UT, USA, pp. 158-167, Mar. 1999.

[7] J. Garcia-Frias and Y. Zhao, "Compression of correlated binary sources using turbo codes", *IEEE Communications Letters*, vol.5, pp. 417-419, Oct. 2001.

[8] A. Aaron and B. Girod, "Compression with side information using turbo codes", *Proc. IEEE Data Compression Conference*, pp. 252–261, Apr. 2002.

[9] A. D. Liveris, Z. Xiong, and C. N. Georghiades, "Compression of binary sources with side information at the decoder using LDPC codes", *IEEE Comm. Letters*, vol. 6, pp. 440–442, Oct. 2002.

[10] N. Gehrig and P. L. Dragotti, "Symmetric and asymmetric Slepian-Wolf codes with systematic and nonsystematic linear codes", *IEEE Commun. Letters,* vol. 9, no. 1, pp. 61-61, Jan. 2005.

[11] P. Tan and P. J. Li, "A general constructive framework to achieve the entire rate region for Slepian-Wolf coding", *Eurasip Signal Processing Journal, Special Issue on Distributed Source Coding*, vol. 86, N°11, Apr. 2006.

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

[13] S. S. Pradhan, J. Chou and K. Ramchandran, "Duality between source coding and channel coding with side information," *IEEE Trans. on Information Theory*, vol. 49, no. 3, pp. 1181-1203, May 2003.

[14] S. Servetto, "Lattice quantization with side information", *Proc. IEEE Data Compression Conference*, Snowbird, UT, USA, Mar. 2000.

[15] Z. Liu, S. Cheng, A. D. Liveris and Z. Xiong "Slepian-Wolf coded nested lattice quantization for Wyner-Ziv coding: High-rate performance analysis and code design", *IEEE Trans. on Information Theory*, vol. 52, no. 10, pp. 4358-4379, Oct. 2006.

[16] Y. Yang, S. Cheng, Z. Xiong and W. Zhao, "Wyner-Ziv coding based on TCQ and LDPC codes", *Proc. Asilomar Conference on Signals, Systems and Computer*, pp. 825-829, Pacific Grove, CA, Nov. 2003.

[17] R. Purit and K. Ramchandran, "PRISM: A new robust video coding architecture based on distributed compression principles", *Proc. Allerton Conf. on Communication, Control and Computing*, Allerton, IL, USA, Oct. 2002.

[18] A. Aaron, R. Zhang and B. Girod, "Wyner-Ziv coding of motion video", *Proc. Asilomar Conf. on Signals, Systems and Computers, Pacific Grove, CA, USA,* Nov. 2002.

[19] A. Aaron, S. Rane and B. Girod, "Wyner-Ziv video coding with hash-based motion-compensation at the receiver", *Proc. IEEE Intl. Conf. on Image Processing, Singapore,* Oct. 2004.

[20] E. Martinian, A. Vetro, J. Ascenso, A. Khisti and D. Malioutov, "Hybrid Distributed Video Coding Using SCA Codes", *Proc. IEEE International Workshop on Multimedia Signal Processing, Victoria, Canada,* Oct. 2006.

[21] K. M. Misra, S. Karande, and Hayder Radha, "Multi-hypothesis based distributed video coding using LDPC codes", *Proc. Allerton Conference on Communication, Control and Computing, Allerton, IL, USA,* Sept. 2005.

[22] M. Tagliasacchi, A. Trapanese, S. Tubaro, J. Ascenso, C. Brites and F. Pereira, "Intra mode decision based on spatio-temporal cues in pixel domain Wyner-Ziv video coding", *Proc. IEEE Intl. Conference on Acoustics, Speech and Signal Processing,* Toulouse, France, May 2006.

[23] M. Maitre, C. Guillemot and L. Morin "3D Model-based frame interpolation for distributed video coding of static scenes", *IEEE Trans. on Image Processing,*, to appear, 2007.

[24] J. Ascenso, C. Brites and F. Pereira, "Improving frame interpolation with spatial motion smoothing for pixel domain distributed video coding", *Proc. 5th EURASIP Conference on Speech and Image*

Processing, Multimedia Communications and Services, Smolenice, Slovak Republic, July 2005.

[25] Z. Li, L. Liu and E. J. Delp, "Wyner-Ziv video coding with universal prediction", *IEEE Trans. on Circuits and Systems for Video Technology,* Vol. 16, no. 11, Pages 1430-1436, Nov. 2006.

[26] P.F.A. Meyer, R.P. Westerlaken, R. K. Gunnewiek and R.L. Lagendijk, "Distributed source coding of video with non-stationary side-information", *Proc. Visual Communications and Image Processing,* Beijing, China, July 2005.

[27] H. Wang, N-M.Cheung, and A. Ortega, "A framework for adaptive scalable video coding using Wyner-Ziv techniques", *Eurasip Journal on Applied Signal Processing,* vol. 2006.

[28] D. Varodayan, A. Aaron and B. Girod, "Rate-adaptive codes for distributed source coding" *Eurasip Signal Processing Journal*, vol. 86, no. 11, pp.3123-3130, Nov. 2006.

[29] Y. Steinberg and N.Merhav, "On successive refinement for the Wyner-Ziv problem", *IEEE Transactions on Information Theory,* vol. 50, no. 8, pp. 1636-1654, Aug. 2004.

[30] Q. Xu and Z. Xiong, "Layered Wyner-Ziv video coding", *IEEE Trans. on Image Processing*, vol. 15, no. 12, pp. 3791-3802, Dec. 2006.

[31] M. Tagliasacchi, A. Majumdar and K. Ramchandran, "A distributed source coding based spatio-temporal scalable video codec", *Proc. Picture Coding Symposium,* San Francisco, CA, USA, Dec. 2004.

[32] Q. Xu, V. Stankovi´c, A. Liveris, and Z. Xiong, "Distributed joint source-channel coding of video", *Proc. IEEE Intl. Conf. on Image Processing*, vol. II, pp. 674-677, Genoa, Italy, Sept. 2005.

[33] S. Shamai, S. Verdu and R. Zamir, "Systematic lossy source/channel coding", , *IEEE Transactions on Information Theory,* vol. 44, no. 2, pp. 564-579, Mar. 1998.

[34] R. Rane, A. Aaron and B. Girod, "Systematic lossy forward error protection for error-resilient digital video broadcasting", *Proc. SPIE Visual Communications and Image Processing, San Jose, CA, USA*, Jan. 2004.

[35] A. Sehgal, A. Jagmohan and N. Ahuja, "Wyner-Ziv coding of video: An error resilient compression framework", *IEEE Trans. on Multimedia,* vol. 6, no. 2, pp. 249-258, Apr. 2004.

[36] M. Flierl and B. Girod, "Coding of multiview image sequences with video sensors," *Proc. IEEE Intl. Conference on Image Processing,* Atlanta, GA, USA, Oct. 2006.

[37] X. Guo, Y. Lu, F. Wu, W. Gao and S. Li, "Distributed multiview video coding" *Proc. of SPIE* vol. 6077, San Jose, CA, USA, Jan. 2006.

[38] M. Ouaret, F. Dufaux and T. Ebrahimi, "Fusion-based multiview distributed video coding" *Proc. ACM International Workshop on Video Surveillance and Sensor Networks,* Santa Barbara, CA, Oct.

2006.

[39] B. Song, O. Bursalioglu, A. Roy-Chowdhury and E. Tuncel, "Towards a multi-terminal video compression algorithm using epipolar geometry," *Proc. IEEE Intl. Conference on Acoustics, Speech and Signal Processing,* PP. II-49 to II-52, Toulouse, France, May 2006.

[40] E. Martinian, A. Behrens, J. Xin and A. Vetro, "View synthesis for multiview video compression," *Picture Coding Symposium,* Beijing, China, Apr. 2006.

Application Scenario	Wireless video cameras	Wireless low-power surveillance	Video conferencing with mobile devices	Mobile video mail	Disposable video cameras
DVC Benefits	Lower encoder complexity and power consumption Flexible allocation of codec complexity Improved error resilience Scalability advantages	Lower encoder complexity and power consumption Lower size and weight Flexible allocation of codec complexity Improved error resilience Scalability advantages Multiview correlation exploitation	Lower encoder complexity and power consumption Flexible allocation of codec complexity Increased resolution for same complexity Improved error resilience	Lower encoder complexity Increased resolution for some power Improved error resilience	Lower encoder complexity and power consumption Small and lightweight devices Flexible allocation of codec complexity
DVC Drawbacks	Higher decoding complexity Lower compression efficiency	Lower compression efficiency Need for a (network) transcoder	Lower compression efficiency Need for a (network) transcoder	Lower compression efficiency No possible encoder playback Need for a (network) transcoder	Lower compression efficiency Higher decoding complexity Absence of a return channel
Application Scenario	**Visual sensor networks**	**Networked camcorders**	**Distributed video streaming**	**Multiview video acquisition**	**Wireless capsule endoscopy**
DVC Benefits	Lower encoder complexity and power consumption Higher coding efficiency compared to current (JPEG) solutions Improved error resilience Scalability advantages Multiview correlation exploitation	Lower encoder complexity and power consumption Small and lightweight devices Improved error resilience Scalability advantages	Flexible allocation of codec complexity Improved resilience and reliability Multiple resolutions handling Scalability advantages, e.g. codec independent scalability	Lower encoder complexity and power consumption Flexible allocation of codec complexity Higher compression efficiency compared to some current (JPEG) solutions Higher image quality for same complexity Multiview correlation exploitation	Lower encoder complexity and power consumption Small and lightweight devices Flexible allocation of codec complexity Scalability advantages, e.g. quality and resolution
DVC Drawbacks	Lower compression efficiency (than alternatives) Higher decoding complexity (than	Lower compression efficiency Higher decoding complexity Early multiview	Lower compression efficiency Higher decoding complexity	Lower compression efficiency Higher decoding complexity Problems with visual occlusions	Lower compression efficiency Higher decoding complexity

Fig. 6. Table 1 - DVC related potential benefits and drawbacks.