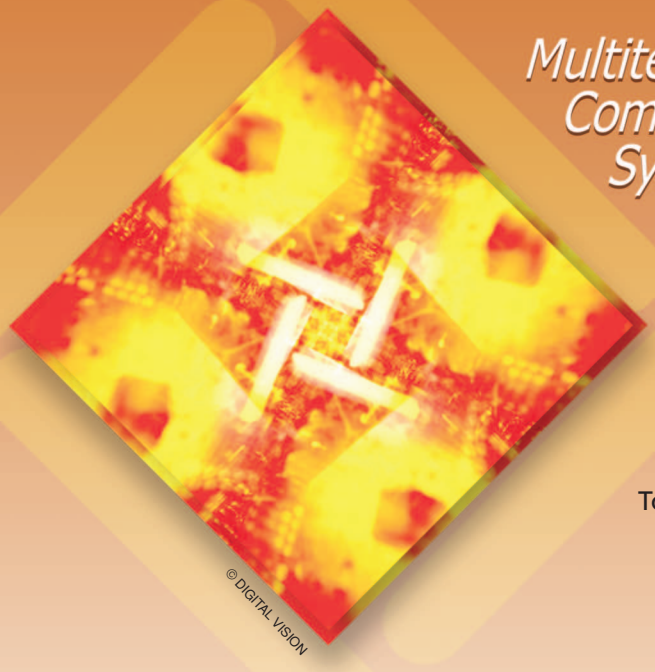


Christine Guillemot, Fernando Pereira, Luis Torres, Touradj Ebrahimi, Riccardo Leonardi, and Jöern Ostermann

Distributed Monoview and Multiview Video Coding

[Basics, problems, and recent advances]

A growing percentage of the world population now uses image and video coding technologies on a 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 is a rather complex exercise. As a consequence, all standard video encoders have a much higher computational complexity than the decoder (typically five to ten 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 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

Digital Object Identifier 10.1109/MSP.2007.904808

different sensors that 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 *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. A comprehensive survey of first DVC solutions can be found in [3]. While, for sake of completeness, basics about DSC are reviewed, this article focuses on DVC latest developments for both monoview and multiview set-ups.

DSC: THEORETICAL BACKGROUND

DSC finds its foundation in the seminal Slepian-Wolf (SW) [4] and Wyner-Ziv (WZ) [11] theorems. Due to space limitation, only the main concepts are recalled. For more details, see [1]–[3].

SLEPIAN-WOLF CODING

Let X and Y be two binary correlated memoryless sources to be losslessly encoded. If the two coders communicate, 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 $R_X + R_Y \geq H(X, Y)$, where $H(X|Y)$ and $H(Y|X)$ denote the conditional entropies between the two sources. Let us consider the particular case where Y is available at the decoder, and has been coded separately at its entropy rate $R_Y = H(Y)$. According to the SW theorem, the source X can be coded losslessly at a rate arbitrarily close to the conditional entropy $H(X|Y)$, if the sequence length tends to infinity. The minimum total rate for the two sources is thus $H(Y) + H(X|Y) = H(X, Y)$. This set-up 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) corresponds to one of the corner points of the SW rate region (see [1]).

The proof of the SW theorem is based on random binning [4], which is nonconstructive, 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]. 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. Channel capacity-achieving codes, block codes [6], turbo codes [7]–[9] or Low Density Parity Check (LDPC) codes [10], 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., a syndrome, or parity bits. The decoder corrects the virtual channel noise, and thus estimates X given the received syndrome or parity bits and the SI Y regarded as a noisy version of the codeword systematic bits.

WYNER-ZIV CODING

In 1976, Wyner and Ziv considered the problem of coding of two correlated sources X and Y, with respect to a fidelity criterion [11]. They have established the rate-distortion (RD) 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 [12] 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.

Practical code constructions based on the WZ theorem thus naturally rely on a quantizer (source code) followed by an SW coder (channel code). The quantizer partitions the continuous-valued source space into 2^{R_s} regions (or quantization cells), where R_s is defined as the source rate in bits/sample. A codeword is associated to each region, thus constructing the source codebook. The SW coder then partitions the source codebook into 2^R cosets, each containing 2^{R_c} (with $R = R_s - R_c$) codewords, and computes the index of the coset containing the source codeword. Only the index I of the coset is transmitted with a transmission rate $R \leq R_s$. The SW decoder recovers the source codeword (or an estimate $\hat{X}_q$ of the quantization index) in a given coset by finding the codeword which is the closest to the observed SI Y. The SW decoder is followed by a Minimum Mean Square Error (MMSE) estimation which searches for $\hat{X}$, the reconstructed value of X, minimizing the expectation $E[(X - \hat{X})^2 | \hat{X}_q, Y]$. A graphical illustration of the WZ coding steps can be found in [6] with the example of scalar quantization. Under ideal Gaussianity assumptions, the WZ limit can be asymptotically achieved with nested lattice quantizers [13], [14].

FROM DSC TO DVC AND POTENTIAL BENEFITS

Video compression solutions today mostly rely on motion-compensated prediction techniques to remove the redundancy between adjacent frames. The encoder searches for the best temporal predictors using motion estimation techniques. It then computes a prediction error which is usually transformed and entropy coded to remove the remaining redundancy in the signal. The motion fields are transmitted and used by the decoder to find the predictors and do the reverse operations. This results in asymmetric systems with a significantly higher encoder complexity due, for a large part, to the motion estimation. This asymmetric structure is well suited for current applications of video compression such as transmission of digital TV,

or video retrieval from servers. However, a large deployment of mobile devices induces the need for a structure with inverted complexity, where video will be encoded using low cost devices and decoded on powerful platforms.

The SW and WZ theorems suggest that, under Gaussianity assumptions, correlated samples of the input video sequence can be quantized and coded independently with minimum loss in terms of RD performance, if they are decoded jointly. This, in principle, implies avoiding the time and energy consuming steps of motion estimation and predictor search in the encoder, with the effect of a complexity shift from coder to decoder as well as increased error resilience. Ideally, only the statistical dependence (or correlation model parameters) between the WZ encoded samples and SI needs to be known to the encoder. However, the application of the WZ principles to video compression requires solving a number of issues which will be discussed in the sequel.

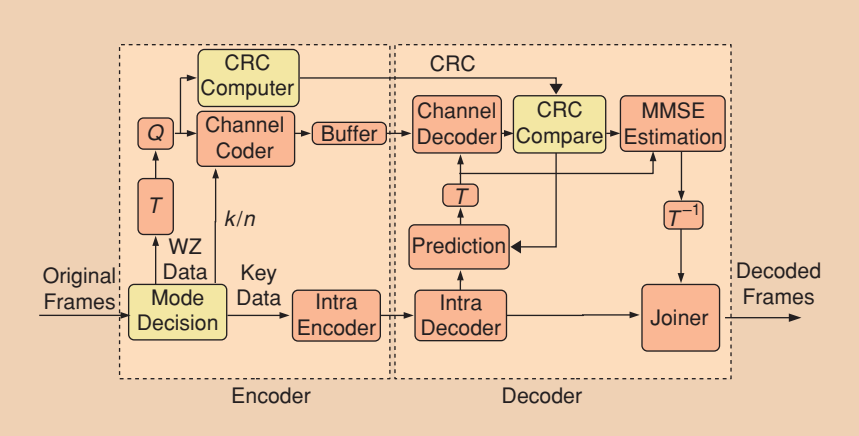
DVC: TOWARDS PRACTICAL SOLUTIONS FOR MONOVIEW SYSTEMS

FIRST DVC ARCHITECTURES

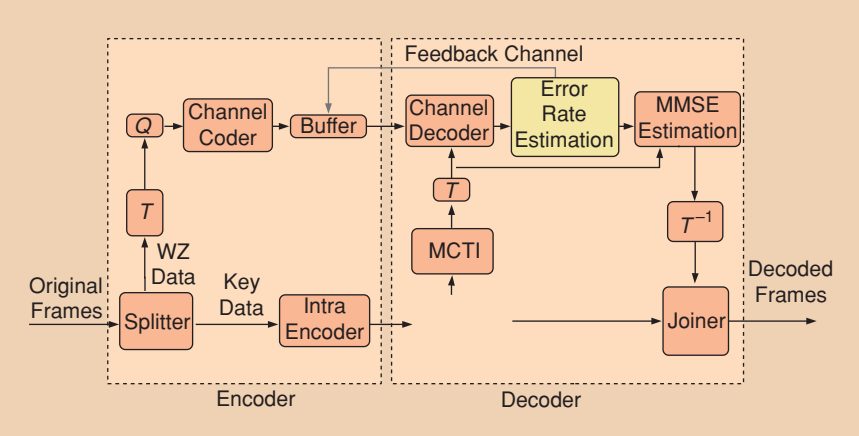
First DVC architectures appeared in 2002 [15], [16]. The WZ principles are applied either in the pixel domain or in the transform domain. Transposing WZ coding from the pixel to transform domain allows us to exploit the spatial redundancy within images, as well as to have correlation models adapted to the different frequency components. A comprehensive overview of the DVC state-of-the-art in 2004 can be found in [3].

A first architecture, called PRISM [15], is depicted in Figure 1. The encoder, based on frame differences, classifies each 16×16 block of the frame into not coded, intracoded, or WZ coded with a set of predefined rates. The rate chosen for a given block depends on the variance of the frame difference which is 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 candidate block.

A second DVC architecture (see Figure 2) has been proposed in [16] in which the WZ coding decision is taken at a frame level. The sequence is thus structured into groups of pictures (GOP), in which selected frames (for example every N frames for a GOP size equal to N), called key frames, are intracoded (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 bit error rate (BER) at the output of the turbo decoder exceeds a given value, more parity or syndrome bits are requested to the encoder via a feedback channel. This allows controlling the bit rate in a more accurate manner

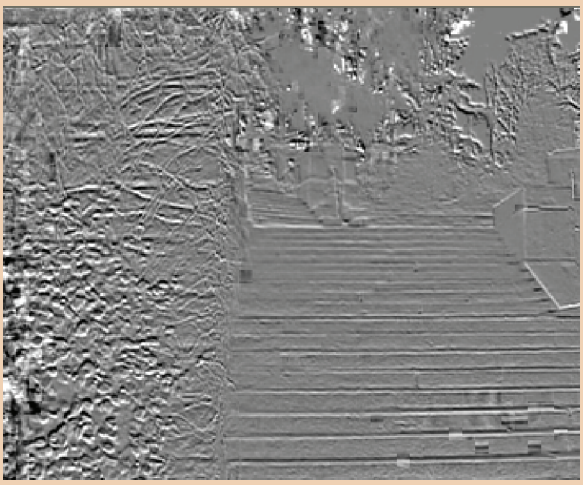


[FIG1] DVC architecture with block-based coding mode selection and rate control at the encoder.

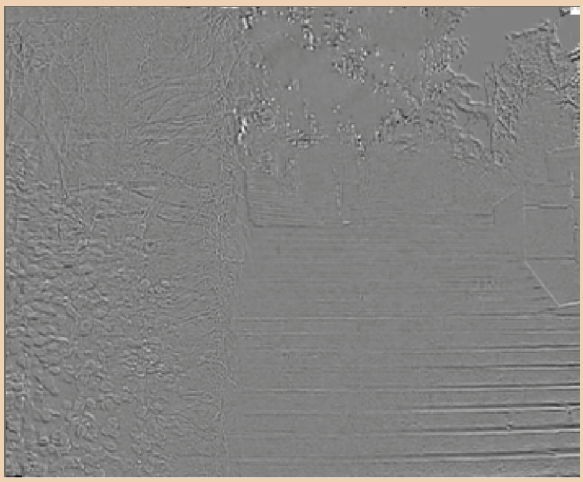


[FIG2] DVC architecture with frame-based coding mode selection and rate control at the decoder.

(a)



(b)



(c)

[FIG3] Correlation noise (difference between interpolated and actual WZ frame) with (a) original sequence, (b) block-based motion-compensated interpolation, and (c) 3-D model and feature points to correct misalignments.

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 two) 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.

OPEN PROBLEMS AND RECENT ADVANCES

RD performances superior to that of H.263+ intraframe coding (a gain of 2 dB for sequences having low motion such as Salesman and Hall monitor [3]) have been reported. However, a significant performance gap relative to H.263+ motion-compensated interframe coding, and H.264/AVC, remains. This gap can be explained by several factors.

SIDE INFORMATION CONSTRUCTION

The decoder must construct SI with minimum distance with the WZ encoded data (i.e., with smallest correlation noise) from previously decoded data. Similarly to predictive coding, it is a problem of reference finding, but this time performed by the decoder. In predictive coding, the encoder searches for the best temporal predictor of the data to encode with block-based motion estimation techniques. The goal is to minimize the error between the predictor and the data to encode (which it knows). In DVC, the decoder must find a predictor (the SI) for data which it does not know.

The first problem is thus to estimate the motion of WZ data (unknown to the decoder) with respect to previously decoded frames. The decoder can only compute motion fields between previously decoded frames which may be distant from one another. An interpolated (or an extrapolated) version of these motion fields, assuming linear motion, is then used to generate a motion field for each WZ frame, which is in turn used for frame interpolation (or extrapolation) to construct the SI. But, the resulting motion fields are unlikely to minimize the distance between SI and WZ data, especially in moving and covered/uncovered regions [see Figure 3(a)]. Slight improvements can be obtained by removing motion discontinuities at the boundaries and outliers in homogeneous regions [17]. Covered/uncovered regions can also be better handled by constructing multiple SI by forward and backward extrapolation rather than by frame interpolation [18].

To help the decoder in its search for the best SI, the encoder can send extra information (CRCs [15] or hash codes [19], [20]), which is some a priori information on the WZ data. The motion-based extrapolation/interpolation step is then embedded in a framework where the decoder has access to multiple candidate SI blocks and checks whether the decoded CRC (or the hash) with each candidate block matches the transmitted CRC. This approach for searching the best SI requires multiple WZ decoding steps, which increases the decoder complexity, and implies transmission rate overhead. Feature points extracted in the WZ frame are transmitted as extra information in [21] to help correcting misalignments in three-dimensional (3-D) model-based frame

interpolation which accounts for scene geometric constraints. In the case of static scenes captured by a moving camera, the approach significantly improves the SI quality (see Figure 3).

CORRELATION MODELLING AND ESTIMATION

The no loss result of the Wyner-Ziv theorem comes under the assumption that the statistical dependence between WZ data and SI is perfectly known to both encoder and decoder, and that it follows a Gaussian distribution. Exact knowledge of the statistical dependence between X and Y is required 1) to characterize the channel in the SW decoder, 2) to perform MMSE estimation in the inverse quantizer, and 3) to help controlling the SW code rate. The RD performance of a DVC system thus strongly depends on its capability to estimate the correlation model parameters.

ALL STANDARD VIDEO ENCODERS HAVE A MUCH HIGHER COMPUTATIONAL COMPLEXITY THAN THE DECODER.

Let us consider the case where WZ coding is performed in a transform domain (as depicted in Figure 1 and 2). The coefficients corresponding to the same frequency information are grouped in a subband, and the correlation parameters are then estimated per subband. Let X denote a WZ sample in a given subband and Y the corresponding SI sample. In practice, the correlation model between X and Y [i.e., the probability density function (pdf) of the difference $Y - X$] is assumed to be Laplacian. In first DVC implementations, the Laplacian parameters were off-line computed for each sequence. A method for online estimation of these parameters at the decoder has been described in [22]. The pdf of the difference $Y - X$ is assumed to match the pdf of the residue (or of its transformed version) between decoded key frames—which the decoder knows—and their motion-compensated versions. The correlation model parameters can then be used to estimate the SW code rate required, and the corresponding value transmitted to the encoder via a feedback channel. This approach implies no processing related to correlation estimation at the encoder, but induces latency and feedback channel usage. Alternatively, depending on latency and/or complexity constraints, the encoder can first estimate the SI which the decoder is likely to have, and then derive the correlation model parameters from the residue $X - \tilde{Y}$ between the WZ data and the SI estimate $\tilde{Y}$ [37]. To avoid increasing too much the encoder complexity, the estimate $\tilde{Y}$ is usually taken as the previously decoded frame (i.e., assuming null motion).

In practice, the samples X are first quantized on K bits $(X_i, i = 1 \ldots K$, where X_i is an independent identically distributed (i.i.d) binary random variable) to be coded bitplane per bitplane with a binary SW code. A bitplane-wise correlation model thus needs to be derived for controlling both the SW coder and decoder. A first model assumes the correlation channel between bitplanes of same significance of WZ and SI data to be binary symmetric, characterized by a crossover probability $p_{co,i} \equiv \Pr(X_i \neq Y_i)$ which varies from bitplane to bitplane. A second model considers the conditional probability $\Pr(X_i|Y, X_{i-1}, ..., X_2, X_1)$. Both probabilities $p_{co,i}$ and $\Pr(X_i|Y, X_{i-1}, ..., X_2, X_1)$ can be easily derived from the Laplacian distribution $\Pr(X|Y)$ or $\Pr(X - Y)$. The crossover probabilities $p_{co,i}$ can also be deduced by measuring, bitplane-wise, the Hamming distance between WZ data and SI estimate $\tilde{Y}$, if the approach retained, e.g., to avoid latency induced by the use of a feedback channel, is to have some estimate of SI at the encoder.

Video signals being highly nonergodic, the correlation channel is in general nonstationary, and the estimation of its parameters may not be accurate. In particular, in regions of occlusions, motion estimation and interpolation are likely to fail, leading to SI with very little correlation with the original data to be WZ coded. This effect can be reflected in the noise model by considering a mixture of Laplacian pdf distributions with higher variance for regions of occlusions [23]. The estimation error is going to impact the SW decoder and MMSE estimation performance, as well as the accuracy of the rate control.

RATE ALLOCATION AND CODING MODE SELECTION

The rate allocation problem involves two aspects: the source code rate control (i.e., the number of quantization levels) and the SW code (i.e. channel coder) rate control. The number of quantization levels is adjusted for a target distortion, assuming perfect SW coding/decoding and targeting a stable PSNR over time for the reconstructed signal. The SW code rate then depends on the correlation between SI and original data.

Let us again consider the case the SW coding is performed bitplane-wise per subband. The rate of the SW code can be estimated from the entropy of the bitplane crossover probability $(H(X_i|Y_i) = -p_{co,i} \log_2 p_{co,i} - (1 - p_{co,i}) \log_2 (1 - p_{co,i}))$. In [25], the entropy of the probability $\Pr(X_i \neq \hat{X}_i)$ averaged over the entire bitplane of a given band, where $\hat{X}_i$ is given by $\hat{X}_i \equiv \arg\max_{b=0,1} \Pr(X_i = b|Y, X_{i-1}, \ldots, X_2, X_1)$, is shown to be a relatively good estimate of the actual rate needed for the SW code. This derivation makes the assumptions that the correlation model is accurate and that the SW code is perfect, which is obviously not the case. This initial rate control can be complemented with a feedback mechanism. If, after sending this initial amount of parity bits, the BER estimated at the output of the SW decoder remains above a given threshold, extra information is requested via a feedback channel. This BER can be estimated from the log likelihood ratios available at the output of the SW decoder [25]. Having an initial rate estimate limits the use of the feedback channel, hence leads to a reduction of delay and decoder complexity. Controlling the rate via a feedback channel requires a rate-adaptive SW code, using e.g., puncturing mechanisms. Syndrome based approaches using punctured LDPC codes are shown to perform poorly because the graph resulting from the puncturing contains unconnected and single-connected nodes [26]. LDPC-based rate-adaptive codes with accumulated syndromes perserving good performance at high compression ratios are described in [26].

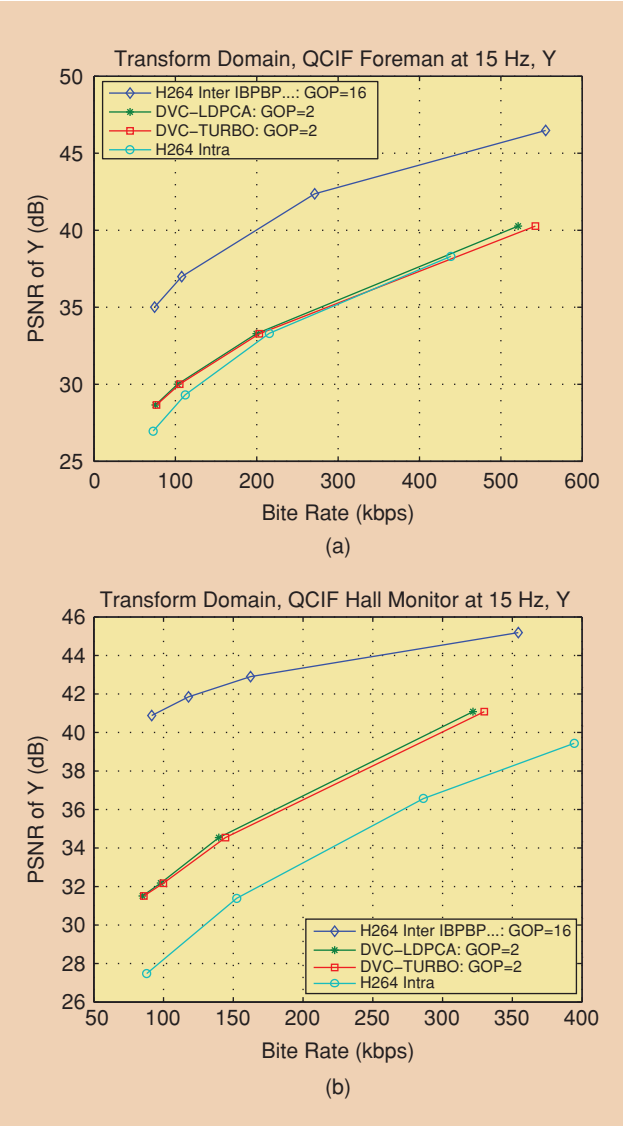
In regions of occlusion, given the low correlation between SI and original data, separate encoding and decoding may outperform WZ coding. As in predictive coding systems, it may thus be beneficial to introduce at the encoder a block-based coding mode (Intra, WZ coded) selection [15], [24]. For deciding the coding mode, the encoder needs to estimate the SI which will be available at the decoder (see previous section). The coding mode selection can be combined with a rate control of the SW code: a rate is thus chosen among a fixed set of possible rates depending on the correlation with the estimated SI [15]. The rate, function of an estimate $\tilde{Y}$ of the SI and not of the actual SI Y available at the decoder, does not match the actual correlation channel.

CAN DSC THEORY AND DVC PRACTICE MEET?

Despite recent advances, DVC RD performance is not yet at the level of predictive coding. The critical steps with respect to RD performance are: 1) finding the *best* SI (or predictor) at the decoder and 2) accurately modeling and estimating the correlation channel. It is shown in [27] that, WZ coding using motion estimation at the encoder for accurate modelling of the displaced frame difference (DFD) statistics and for signalling the best SI to the decoder, give performances close to those of predictive coding. However, this comes at the cost of an encoder complexity comparable to the one in predictive coding systems.

The suboptimality of these two steps shifted to the decoder depends on the motion characteristics of the video sequence. Fast motion negatively impacts the SI quality. Figure 4 illustrates the performance gap between a DVC architecture based on the feedback channel (as depicted in Figure 2) with punctured turbo and LDPC codes [26] for two sequences (with fast and low motion) at 15 Hz. For sequences with low motion or higher frame rates (e.g., 30 Hz), a RD performance gain close to 3 dB is achieved, with a significantly lower complexity, compared with H.264/AVC Intra. With fast motion or low frame rates, this is not always the case. Dynamic GOP size adaptation at the encoder, according to sequence motion characteristics, further improves the RD performance, however at the expense of increased encoding complexity.

The DVC paradigm brings flexibility for shifting part of the encoder complexity to the decoder, i.e., for coder/decoder complexity balancing. Low encoding complexity constraints have so far been central to the design of first DVC solutions. But, the various trade-offs between RD performance and coder/decoder complexity balancing, according to applications constraints, are not yet fully understood and remain to be explored. Beyond the complexity-performance trade-off advantage, the DVC paradigm presents interesting features in terms of error resilience and for scalable coding.



[FIG4] PSNR-rate performance of H.264/AVC Intra, H.264/AVC Inter, DVC with punctured turbo codes, DVC with punctured LDPCA codes.

WZ CODING FOR ERROR-RESILIENT VIDEO TRANSMISSION

Predictive video coding is very sensitive to channel errors: Bit errors or packet losses lead to predictive mismatch, also known as the drift effect, which may result in a significant quality degradation of the reconstructed signal. Predictive decoders, when used in noisy transmission environments, are followed by a post-processing step known as error concealment to limit the catastrophic effect of drift and error propagation. The reconstructed signal remains however significantly impaired.

In DVC, in presence of errors, the SI quality is also going to degrade, resulting, similarly to predictive coding, into a drift effect or predictive mismatch at the decoder. The SI Y can only be constructed from concealed data, and will be denoted $\tilde{Y}$. The virtual channel for the WZ coding problem is then defined by the distribution of $X - \tilde{Y}$ instead of $X - Y$. The corresponding errors will be corrected if they remain within the power of correction of the SW code, which then operates as a joint source-channel code. The rate of the SW coder can thus be set in order to correct the noise of the degraded correlation channel [28]. Note that architectures in which the decoder searches—with methods close to motion estimation—for the best SI are more amenable to reduce the noise of the degraded correlation channel.

Alternatively, WZ coding can be used as a systematic lossy forward error correction (FEC) technique. Extra information is sent on an auxiliary channel to mitigate the drift effect. This idea has been initially suggested in [29] 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 [30]–[32] to the problem of robust video transmission. The video sequence is first conventionally encoded, e.g., using an MPEG coder. The sequence is also WZ encoded. In case of errors, once the conventional bitstream is decoded, error concealment techniques are applied to construct the SI used to decode the WZ stream. In [32], for some frames called peg frames, the indexes of the coset to which belong the symbols of a given image I_p are transmitted in addition to the residue of the temporal prediction performed by the conventional coder. The error propagation due to the drift effect is thus confined between two peg frames. In [30], the correlation noise of the global channel (correlation plus transmission-induced SI distortion) is modelled, and a subset of transform coefficients of the conventional stream is WZ coded by the auxiliary coder. In the above approaches, the predictively encoded bitstream constitutes the systematic part of the information which can be protected with classical FEC. The WZ encoded stream is an extra coarser description of the video sequence, and is redundant if there is no transmission error. This can be regarded as unbalanced multiple description coding.

Avoiding the cliff effect of conventional FEC, systematic lossy error protection based on WZ coding has been shown to lead to a more graceful degradation of the reconstructed video quality [33]. However, the research in the area of WZ coding based robust video transmission is still at the level of preliminary proofs of concepts. A mature solution with precise assessment of its error resilience benefits under realistic communication scenarios and against conventional FEC is still missing. How to estimate the channel parameters (which has also to account for the distortion induced on the SI by the transmission noise), and control the rate of the codes accordingly, also remain open issues.

LAYERED WZ CODING

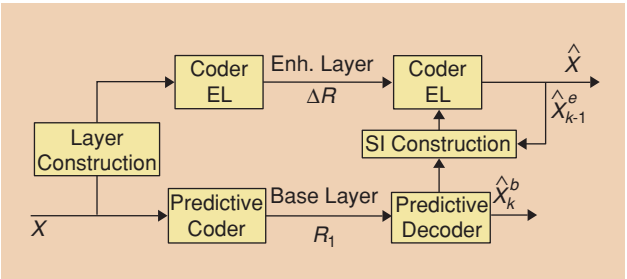
Scalable video coding (SVC) is attractive for applications such as streaming on heterogeneous networks and/or towards terminals having different resolutions and capabilities. SVC solutions are often based on layered signal representations including closed-loop inter-layer prediction. The problem of layered predictive coding, similarly as temporal predictive coding, can be re-cast into a problem of distributed source coding, with similar features in terms of coder/decoder load balancing and error resilience. While in layered coding, the refinement signals are computed from coded and decoded realizations of lower layers, with WZ coding, only the correlation model between WZ data (within one layer) and SI reconstructed from lower layers needs to be known.

The encoding of the refinement signals becomes to some extent independent of the codec used in lower layers, the only constraint being that the correlation noise between the SI reconstructed from lower layers and the WZ data is within the power of correction of the SW code. Theoretic conditions so that successive refinement in a WZ setting can asymptotically achieve the WZ RD function in each layer, i.e., so that $R_1 + \Delta R = R^*_{X|Y_2}(D_2)$, where $R_1 = R^*_{X|Y_1}(D_1)$ is the WZ bound for a layer 1, have been formulated in [34]. In practical systems, this condition which assumes that SI Y_2 in layer 2 does not bring extra information to the one used in layer 1 is rarely verified. SI constructed from previously decoded frames on the enhancement layer is likely to bring extra information to the one used on the base layer.

Let $\hat{X}^b_k$ and $\hat{X}^e_k$ denote the decoded base and enhancement layers for frame k. Let $\hat{X}^{e,j}_k$, $j = 1, \ldots l-1$ be the l first decoded bitplanes of $\hat{X}^e_k$. A SNR scalable scheme is proposed in [35] where the base layer uses a standard codec, and bit planes of the enhancement layers are WZ encoded (as shown in Figure 5). The image reconstructed from decoded base $\hat{X}^b_k$ and enhancement layers $\hat{X}^j_k$, $j = 1, \ldots l-1$ is used as SI to decode X^l_k. The temporal redundancy in enhancement layers is not exploited. In [36], a spatial and temporal scalable codec based on PRISM is described. For spatial scalability, motion vectors estimated in the conventional base layer codec are used to choose between spatial, temporal and spatio-temporal prediction, as well as between correlation parameters (trained off-line) for each type of predictor. For the temporal scalability, higher layer motion vectors are inferred from those of the base layer. In [37], $\hat{X}^b_k$ is used to compute a residue $U_k = X_k - \hat{X}^b_k$ using closed-loop inter-layer prediction. This residue is then either coded with entropy source codes or WZ coded using $V_k = \tilde{X}^e_k$ as SI, depending on whether the temporal correlation is low or high.

MULTIVIEW DISTRIBUTED VIDEO COMPRESSION

Storage and transmission of multiview video sequences of the same scene involve large volumes of redundant data. These data can be efficiently compressed with techniques which compress the signals jointly, exploiting correlation in the temporal direction as well as correlation between views. Techniques compensating the displacement of an object from one view to the other, called disparity, are used to remove inter-view correlation. Disparity vectors are function of depth, i.e., of the focal length and positions of the cameras. These techniques are referred to as—pixel-based or block-based—disparity-compensated view prediction techniques. Prediction



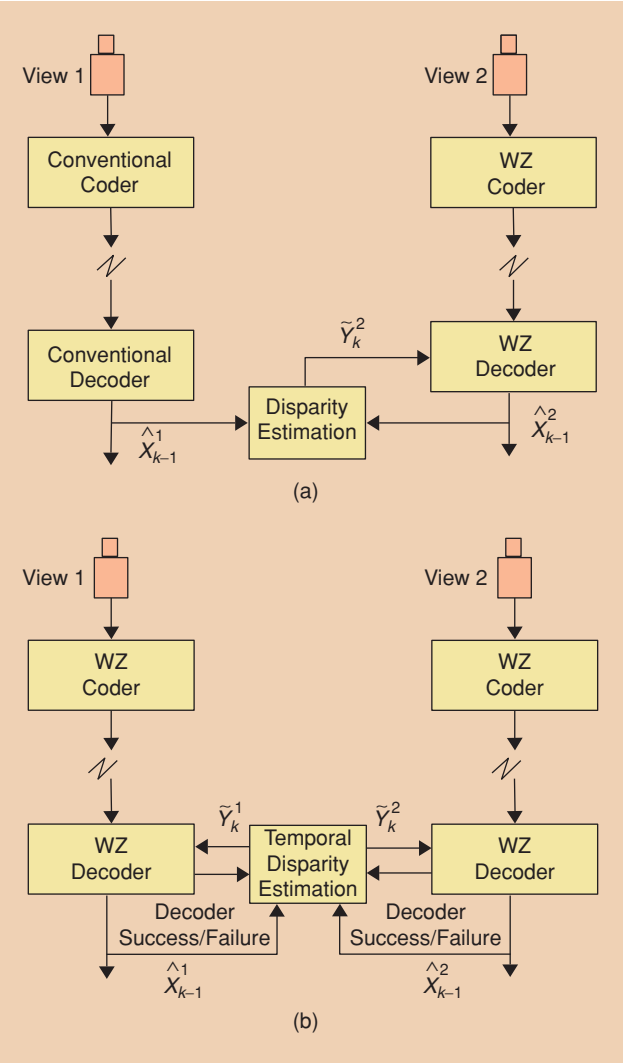
[FIG5] Layered WZ coding/decoding structure with predictive coder in the base layer.

techniques based on a synthesis of intermediate views are alternatives to disparity-compensated techniques [38]. These approaches are however more complex as they require estimating depth maps and constructing 3-D models of the scene.

The promise of DVC is to allow exploiting correlation between views without—or with limited—inter-sensor (that is inter-camera) communication, for infrastructures with limited bandwidth and power consumption contraints. The problem of distributed multiview coding has been first addressed for arrays of video cameras capturing static scenes and for light fields. Here, we concentrate on distributed multiview video compression. Same questions related to SI construction (or prediction) at the decoder side, and on correlation modelling and estimation, as in monoview systems, arise.

VIDEO COMPRESSION SOLUTIONS TODAY MOSTLY RELY ON MOTION-COMPENSATED PREDICTION TECHNIQUES TO REMOVE THE REDUNDANCY BETWEEN ADJACENT FRAMES.

However, in addition to temporal and/or inter-layer dependencies, in multiview DVC, the SI has to also account for inter-view dependencies. Capturing inter-view dependency turns out to be more difficult than for temporal dependencies, as, in general, multi-view images contain disparities much larger than displacements between successive frames. The source of occlusions also differs: in multi-view, occlusion occurs when part of the scene can only be observed by one of the cameras due to depth discontinuities or finite viewing, while, in monoview, occlusion results from objects motion. Several set-ups (two examples are depicted in Figure 6) with different implications on sensor nodes communication and coding/decoding complexity, have been considered. The RD performance gap between joint and distributed compression of multi-view sequences remains large.



[FIG6] SI estimation: (a) disparity estimation based on previously decoded frames; (b) switch between temporal and disparity-based inter-view SI.

INTER-VIEW SIDE INFORMATION

In joint coding systems, disparity vector fields are estimated by the encoders, in order to find the best inter-view predictors. The disparity vector fields are perfectly known to all encoders of the multi-view sequences. In a DVC scenario with no inter-camera exchange, disparity estimation must be performed by the decoder. One approach is to use, for the current frames of the multiview sequences, disparity vector fields estimated on previously decoded frames [39]. The disparity-based SI, estimated from previously decoded frames at time $k-1$, is used to WZ decode the frame at instant k. The resulting uncertainty on the disparity vector fields translates into a rate loss for distributed coding. In [39], it is however shown that decoding with disparity-compensated SI reduces the bit rate by up to 10% over decoding without SI.

Block-based disparity compensation is only applicable in the case of rectified views on a co-linear line, with a viewing axis perpendicular to the baseline. Alternatively, and provided that the scene can be approximated by a planar surface (i.e., all objects lie on a plane), that the scene is static or that the camera motion is a pure rotation around its optical center, disparities can be better represented by global models instead of simple block-based translational models. An eight-parameter homography is used in [40]. The homography is a 3×3 matrix that relates one view to another in the homogenous coordinates system.

The disparity between corresponding points in different views depends on camera positions and scene geometry. The disparity search can thus be constrained on the epipolar geometry: given a point in one view, its corresponding point in the other view lies on the epipolar line. The epipolar constraint is actually used to reduce the search of correspondences to a one-dimensional (1-D) problem [41]. Motion vectors are estimated on each view of the stereo set-up and exchanged between sensors. Together with epipolar constraints the motion vectors help the disparity search. However, the complexity of each sensor node

which supports the motion estimation constrained along the epipole line remains high. A disparity search constrained along the epipolar line is also performed in each view of the multiview set-up of [42] and depicted in Figure 6(b).

FUSION OF TEMPORAL AND INTER-VIEW SIDE INFORMATION

In [39] and [41], the temporal correlation is exploited at the encoder using classical techniques such as a motion-compensated wavelet transform and predictive coding respectively. In [39], one view is considered as the reference view encoded with a motion-compensated temporal filtering approach. The other views are also first temporally transformed with a motion-compensated wavelet transform. Each temporal subband is WZ coded using inter-view disparity-compensated SI. The encoder on each sensor node then remains rather complex.

In wireless scenarios, with constraints of low-power consumption, distributed compression may be preferable to predictive coding also along the temporal direction, in which case both temporal and inter-view SI need to be constructed. Depending on cameras positions, on spatial and temporal resolutions of the sequences, on motion in the scene, temporal correlation may be higher than inter-view correlation or vice-versa. A switch between temporal and disparity-based SI is used in [42]. All the views are first decoded using temporal SI. If decoding fails for a particular block, then disparity search is performed on the available reconstructions [see Figure 6(b)]. These two types of SI also lead to increased error-resilience. In [40], considering a particular set-up in which some views are intracoded while others are encoded with a structure as shown in Figure 2, including both intracoded and WZ-coded frames, a fusion is done between temporal SI constructed by interpolation of key frames and homography-based inter-view SI. The decision mask is estimated from the best prediction on temporally adjacent key frames. Preliminary 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.

CONCLUDING REMARKS

Compared with predictive coding, DVC holds promises for a number of applications: a more flexible coder/decoder complexity balancing, increased error resilience, and the capability to exploit inter-view correlation, with limited inter-camera communication, in multiview set-ups. DVC shows benefits in layered representations, with increased error resilience, and to some extent, independence between codecs used in the different layers. However, despite the growing number of research contributions in the past, key questions remain to bring monoview and multiview DVC to a level of maturity closer to predictive coding: estimating at encoder or decoder the virtual correlation channel from unknown—or only partially known—data; finding the best SI at the decoder for data not—or only partially—known. Solutions to the above questions have various implications on coder/decoder complexity balancing, on delay and communication topology (e.g., need for a feedback channel), and RD performance. These various trade-offs, the RD performance limits versus application constraints in terms of delay, coder/decoder complexity trade-offs, precise error resilience benefits under realistic communication scenarios, remain to be carefully addressed for real application take-up.

ACKNOWLEDGMENTS

The authors would like to thank the European Commission for its support in Distributed Video Coding research done in the context of the DISCOVER project of the IST FP6 program (htpp://www.discoverdvc.org). The authors would also like to thank P. Correia (IST), E. Acosta (UPC), X. Artigas (UPC), M. Ouaret (EPFL), F. Dufaux (EPFL), D. Kubasov (INRIA), K. Lajnef (INRIA), M. Dalai (UNIBS), and S. Klomp (UH) for their contributions to this paper.

AUTHORS

Christine Guillemot (Christine.Guillemot@irisa.fr) is currently Director of Research at INRIA. She holds a PhD degree from ENST (Ecole Nationale Superieure des Telecommunications) Paris. Her research interests are signal and image processing, video coding, and joint source and channel coding for video transmission over the Internet and over wireless networks. She has served as associated editor for IEEE Transactions on Image Processing and for IEEE Transactions on Circuits and Systems for Video Technology. She is a member of the IMDSP and MMSP technical committees.

Fernando Pereira (fp@lx.it.pt) is currently Professor at Instituto Superior Técnico. He is an Area Editor of the *Signal Processing: Image Communication Journal* and is or has been an Associate Editor of *IEEE Transactions of Circuits and Systems for Video Technology*, *IEEE Transactions on Image Processing*, and *IEEE Transactions on Multimedia*. He also served as IEEE Distinguished Lecturer. He has contributed more than 150 papers on various areas of video processing, notably video analysis, processing, coding and description, and interactive multimedia services.

Luis Torres (luis@gps.tsc.upc.edu) is currently a Professor at the Technical University of Catalonia, Barcelona, Spain. His main interests are image and video coding, image and video analysis, and face recognition. He has published more than 120 papers in journals and conferences. Luis Torres is a Senior Member of the IEEE. He served as Associate Editor, IEEE Transactions on Image Processing, and as General Chair, International Conference on Image Processing (ICIP 2003). He is currently Associated Editor for the Electronic Journal of Imaging of the SPIE.

Touradj Ebrahimi (Touradj.Ebrahimi@epfl.ch) is a Professor of multimedia signal processing at EPFL, where he is involved in teaching and research in visual information compression (image, video, 3D), information processing (image and video segmentation and tracking, multimedia content quality metrics, feature extraction and tracking), and multimedia security (watermarking, conditional access, data integrity, data hiding).

He was the recipient of the best paper in IEEE Transactions on Consumer Electronics in 2000, and 4 ISO certificates for contributions to MPEG and JPEG standards. He has been author or co-author of over 150 papers, and holds 15 patents.

Riccardo Leonardi (riccardo.leonardi@ing.unibs.it) has obtained his Diploma (1984) and Ph.D. (1987) degrees in Electrical Engineering from the Ecole Polytechnique Fédérale de Lausanne. After a few years spent at AT&T Bell Laboratories (NJ, USA) he was appointed Chair at the University of Brescia to set up the research activities in Communications and Signal Processing. His main research interests cover visual communications, and content analysis of multimedia information. He has published more than 100 papers in these fields. Prof. Leonardi has also established all undergraduate and graduate degrees granted by its University in the field of Telecommunication.

Jöern Ostermann (ostermann@tnt.uni-hannover.de) is Full Professor and Head of the Institute für Informationsverarbeitun at the Leibniz Universität Hannover, Germany. In 1998, he received the AT&T Standards Recognition Award and the ISO award. He is a Fellow of the IEEE. Jöern served as a Distinguished Lecturer of the IEEE CAS Society. He published more than 50 research papers and book chapters. He is co-author of a graduate level text book on video communications. He holds 22 patents.

REFERENCES

[1] S.S. Pradhan, J. Kusuma, and K. Ramchandran, "Distributed compression in a dense micro-sensor network," *IEEE Signal Processing Mag.*, 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 Mag.*, vol. 21, pp. 80–94, Sept. 2004.

[3] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proc. IEEE (Special Issue on Advances in Video Coding and Delivery)*, vol. 93, no. 1, pp. 71–83, Jan. 2005.

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

[5] A. Wyner, "Recent results in the Shannon theory," *IEEE Trans. Inform. 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," *in Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 1999, pp. 158–167.

[7] J. Bajcsy and P. Mitran "Coding for the Slepian-Wolf problem with turbo codes," *in Proc. Global Commun. Symp.*, San Antonio, TX, Nov. 2001, pp. 1400–1404.

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

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

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

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

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

[13] S. Servetto, "Lattice quantization with side information," *in Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 2000, pp. 510–519.

[14] X. Wang and M.T. Orchard, "Design of trellis codes for source coding with side information at the decoder," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 2001, pp. 361–370.

[15] R. Purit and K. Ramchandran, "PRISM: A new robust video coding architecture based on distributed compression principles," in *Proc. Allerton Conf. Commun., Control Comput.*, Allerton, IL, Oct. 2002.

[16] A. Aaron, R. Zhang, and B. Girod, "Wyner-Ziv coding of motion video," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, Nov. 2002.

[17] J. Ascenso, C. Brites, and F. Pereira, "Improving frame interpolation with spatial motion smoothing for pixel domain distributed video coding," in *Proc. 5th EURASIP Conf. Speech Image Processing, Multimedia Commun. Services*, Smolenice, Slovak Republic, July 2005.

[18] K.M. Misra, S. Karande, and H. Radha, "Multi-hypothesis based distributed video coding using LDPC codes," in *Proc. Allerton Conf. Commun., Control Comput.*, Allerton, IL, Sept. 2005.

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

[20] E. Martinian, A. Vetro, J. Ascenso, A. Khisti, and D. Malioutov, "Hybrid distributed video coding using SCA codes," in *Proc. IEEE Int. Workshop Multimedia Signal Processing*, Victoria, Canada, Oct. 2006, pp. 258–261.

[21] M. Maitre, C. Guillemot, and L. Morin, "3-D Model-based frame interpolation for distributed video coding of static scenes," *IEEE Trans. Image Processing*, vol. 16, no. 5, pp. 1246–1257, May 2007.

[22] C. Brites, J. Ascenso, and F. Pereira, "Studying temporal correlation modeling for pixel based Wyner-Ziv video coding," in *Proc. IEEE Int. Conf. Image Processing*, Oct. 2006, pp. 273–276.

[23] R.P. Westerlaken, R.K. Gunnewiek, and R.L. Lagendijk, "The role of the virtual channel in distributed source coding of video," in *Proc. IEEE Int. Conf. Image Processing,* Genoa, Italy, Sept. 2005, vol.1 , pp. 581–584.

[24] 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," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Processing*, Toulouse, France, May 2006, vol. 2, pp. 57-60.

[25] D. Kubasov, K. Lajnef, and C. Guillemot, "A hybrid encoder/decoder rate control for a Wyner-Ziv video codec with a feedback channel," in *Proc. IEEE Multimedia Signal Processing Workshop, MMSP*, Chania, Crete, Greece, Oct. 2007.

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

[27] A. Sehgal, A. Jagmohan, and N. Ahuja, "Scalable video coding using Wyner-Ziv codes," in *Proc. Picture Coding Symp.*, San Francisco, CA, Dec. 2004.

[28] Q. Xu, V. Stanković, A. Liveris, and Z. Xiong, "Distributed joint source-channel coding of video," in *Proc. IEEE Int. Conf. Image Processing*, Genoa, Italy, Sept. 2005, vol. II, pp. 674–677.

[29] S. Shamai, S. Verdu, and R. Zamir, "Systematic lossy source/channel coding," *IEEE Trans. Inform. Theory*, vol. 44, no. 2, pp. 564–579, Mar. 1998.

[30] J. Wang, A. Majumdar, K. Ramchandran, and H. Garudadri, "Robust video transmission over a lossy network using a distributed source coded auxiliary channel," in *Proc. Picture Coding Symp.,* San Francisco, CA, Dec. 2004.

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

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

[33] A. Majumdar, R. Purit, P. Ishwar, and K. Ramchandran, "Complexity/performance trade-offs for robust distributed video coding," in *Proc. IEEE Int. Conf. Image Processing*, vol. II, Genoa, Italy, Sept. 2005, pp. 678–681.

[34] Y. Steinberg and N. Merhav, "On successive refinement for the Wyner-Ziv problem," *IEEE Trans. Inform. Theory*, vol. 50, no. 8, pp. 1636–1654, Aug. 2004.

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

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

[37] H. Wang, N.-M. Cheung, and A. Ortega, "A framework for adaptive scalable video coding using Wyner-Ziv techniques," *Eurasip J. Appl. Signal Processing*, vol. 2006, Article ID 60971, 18 pages.

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

[39] M. Flierl and B. Girod, "Coding of multiview image sequences with video sensors," in *Proc. IEEE Int. Conf. Image Processing*, Atlanta, GA, Oct. 2006, pp. 609–612.

[40] M. Ouaret, F. Dufaux, and T. Ebrahimi, "Fusion-based multiview distributed video coding," in *Proc. ACM Int. Workshop Video Surveillance Sensor Networks*, Santa Barbara, CA, Oct. 2006, pp. 139–144.

[41] B. Song, O. Bursalioglu, A. Roy-Chowdhury, and E. Tuncel, "Towards a multi-terminal video compression algorithm using epipolar geometry," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Processing*, Toulouse, France, May 2006, pp. II-49–II-52.

[42] C. Yeo and K. Ramchandran, "Distributed video compression for wireless camera networks," in *Proc. SPIE Conf. Video Image Commun., VCIP 2007*, San Jose, CA, Feb. 2007.

SP