Distributed video codec with complexity scalability based on mixed resolution coding

Bruno Macchiavello*, Ricardo L. de Queiroz**
*Department of Computer Science
**Department of Electrical Engineering
Universidade de Brasilia, Brazil
Emails: {bruno,queiroz}@image.unb.br

Debargha Mukherjee
Hewlett Packard Laboratories
Hewlett Packard
Palo Alto, USA
Email: debargha.mukherjee@hp.com

***Abstract*—Recently, new requirements in video coding have emerged. These requirements include bandwidth fluctuation, quality of service and energy constraints. Typically, the encoder has a higher complexity, requiring more energy consumption than the decoder. A new paradigm in video coding, called distributed video coding, enables a reversed complexity coding mode, where the decoder requires more computational effort than the encoder. Distributed video coding is based on the Wyner-Ziv theorem for separately coding but jointly decoding correlated sources. A distributed video codec can be used in power constrained scenarios, like in hand-held mobile devices. We propose a mixed resolution framework based on full resolution reference frames and spatial-reduction-based coding of the non-reference frames. Improved rate-distortion performance is achieved by enabling better side-information generation at the decoder and better rate-allocation at the encoder. The framework enables reduced encoding complexity by low-resolution encoding of the non-reference frames, followed by Wyner-Ziv coding of the residue. The quantized transform coefficients of the residual frame are mapped to cosets without the use of a feedback channel. For that purpose, a study to select optimal coding parameters in the creation of the memoryless cosets is made. Furthermore, a correlation estimation mechanism that guides the parameter choice process is proposed. At the decoder, coset decoding is carried using the side-information to obtain a higher quality version of the decoded frame. Results for the coding complexity and rate-distortion performance are presented using the H.264/AVC codec. It is shown that the proposed Wyner-Ziv coding mode is competitive. It allows a scalable complexity reduction and supports a low-decoding-complexity mode.**

***Keywords*-Distributed video coding; Wyner-Ziv; H.264; scalable complexity.**

I. Introduction

Digital video compression is necessary to perform an efficient coding, either for storage or transmission of the video signal. video compression can be defined as the reduction of the data required to encode a sequence of frames. Currently, most video coding standards are designed for broadcast-type applications, where there is a single more powerful encoder and various low-complexity decoders. Therefore, the encoder requires much more computational effort than the decoder [1], mainly due to mode search and motion estimation in finding the best inter coding predictor.

However, new requirements in video coding have emerged. Requirements such as bandwidth fluctuation, energy constraints and quality of service (QoS) can get to be as important as the compression rate. The need for a scalable video codec has increased. There are codecs with SNR (signal to noise ratio) scalability for the H.264/AVC standard [2]–[4]. This extension is known as Scalable Video Coder (SVC), and can achieve some of the new coding requirements. Other coders with SNR scalability were previously presented [5]–[7], and there is also a previous work in order to obtain and evaluate the theoretical limits of the rate-distortion (RD) performance for compression algorithms of video with adaptive bit-rate [8].

Distributed source coding (DSC), has its roots in the theory of coding correlated sources developed by Slepian and Wolf [9] for the lossless case and Wyner and Ziv [10] for the lossy case. It has recently become the focus of different kinds of video coding schemes [11]–[19]. As mention before, digital video standards have high-complex encoders [1]. On the other hand, the decoder complexity is low following a broadcast-oriented model. In scenarios where real-time encoding is required within a limited power environment, it may be useful to consider the distributed video coding (DVC) paradigm. DVC operates with reverse complexity, which means that the encoder complexity is shifted to he decoder. DVC fits the scenario for power-constrained (mobile, hand-held) devices.

In realistic video communications using mobile power-constrained devices, it is important to make some observations: (i) It may not necessary for the video encoder to always operate in a reversed complexity mode, the DVC mode may be turned on only when available battery power drops; (ii) the complexity reduction should not be achieved at a substantial cost in bandwidth; (iii) the decoder must support a mode of operation where at least a low quality version can be decoded immediately with low complexity, since the video communicated from one mobile device may be received and played back in real-time at another mobile device; and (iv) while much of the prior work in DVC [12], [18], [20] require a feedback channel, it is more practical to consider blind case, where the decoder is not necessarily required to attempt distributed decoding immediately after reception. A framework without feedback channel allows DVC coding across all power-constrained devices. However, it increases the difficulty of choosing the correct coding parameters in order to approach the Slepian-Wolf frontier.

Much of the prior work [12] addressing the reversed complexity scheme, use periodic key frames, coded in INTRA (I) mode, with multiple motion-free Wyner-Ziv (WZ) frames in between. Then, without motion estimation, the encoder is substantially less complex than the decoder. However, this framework limits the RD efficiency. One obvious way to improve compression efficiency is to replace key frames (I) with conventional predictive P- or B- frames. However, the quality of the side-information (SI) would still limit the performance. In order to substantially improve the quality of the SI generated at the decoder, a variation of this class of methods [13], [21] transmits an auxiliary information for each block coded to help the block matching process. Nevertheless, the rate needed for the hash is often prohibitive. Another approach [22] uses a highly compressed version of each WZ

Note: This work relates to a Doctoral thesis

frame, with zero-motion vectors, as a frame-hash to improve the SI quality. However, the performance of all these methods, in the no-feedback-channel case, is heavily dependent on the accuracy of the rate-allocation mechanism. Some studies have addressed the rate allocation problem for the motion-free WZ architecture when a feedback channel is not employed [23]–[25]. Normally, a simple difference between pixel values of the original frame and a reference frame is used to predict the bit error probabilities.

An alternative approach is to use spatial-reduction based DVC [26]–[28]. Spatial-reduction can achieve lower complexity coding even in conventional codecs [29]. In the particular case [26] that we explore in this paper, the key frames are conventional I, P or reference-B frames coded at full resolution. The intermediate WZ frames are coded in two layers: a base layer coded at reduced resolution followed by an enhancement layer that uses WZ coding of the Laplacian residue. Such a mixed resolution framework [28] can be implemented as an optional coding mode in any existing video codec standards [1], [30]. The encoding complexity is reduced through lower resolution encoding of the intermediate WZ frames. At the decoder, if only the low resolution (LR) base layer is decoded we will have achieved low-complexity decoding at real-time. For decoding the enhancement layer, we propose a motion-based multi-hypothesis SI generation method [31]–[33]. Other works have used iterative SI generation techniques [34]–[36]. All of them assume key frames are intra coded and the intermediate frames are entirely WZ coded. A method for estimating the correlation statistics is also proposed [37], [38]. An optimal parameter choice mechanism in the creation of the memoryless cosets [28], [38] is presented. It has been shown [39], that for such decoding there should not be substantial improvement if more complex channel codes were used.

Note that, DVC with spatio-temporal scalability has been previously addressed [16]. The use of spatial scalability for coding the key frames may degrade the RD performance at full spatial resolution. Also, layered Wyner-Ziv codecs with similar results to the MPEG-4/H.26L FGS has been proposed [14], [15] based on recent theoretical results on successive refinement of Gaussian sources for WZ coding [40]. However, the framework of this work is different from the previous works.

This paper is organized as follows. The framework is described in Section II. The generation of the SI is introduced in Section III, in Section IV we show how the parameters for the coset generation are chosen based on the estimated variances. We describe how the variances are estimated in Section V. In section VI we discuss some modifications to the entropy coder of the H.264/AVC, in order ti improve its performance for the WZ layer. Finally, the implementation results of the proposed method using the H.264/AVC codec are presented in Section VII, followed by the conclusions in Section VIII.

II. Mixed Resolution Distributed Video Codec

The proposed complexity scalable framework is based on non-reference WZ frames (NRWZ) [26], [28]. The complexity reduction is applied only to these NRWZ frames, which are coded at low spatial resolution. On the other hand, the reference frames are coded exactly at full resolution. The LR base layer if formed by the NRWZ frames coded at LR and the reference frames, the enhancement layer if formed by the high frequency content of the NRWZ frames. If the receiver decodes only the base layer, it can immediately playback a lower quality version of the video with low complexity, leaving full decoding of the NRWZ frames to offline processing. The framework does not limit the number of NRWZ frames between the reference frames. Ideally, the number of NRWZ frames can be varied dynamically based on the complexity reduction required and on the target quality. Figure 1 depicts typical frame configurations. In Fig. 1(a), the B frames of a conventionally coded sequence have been converted into NRWZ B-like frames, while Fig. 1(b) shows a similar sequence with lower complexity. Fig. 1(c) shows a low delay case where NRWZ frames are used like P-frames.

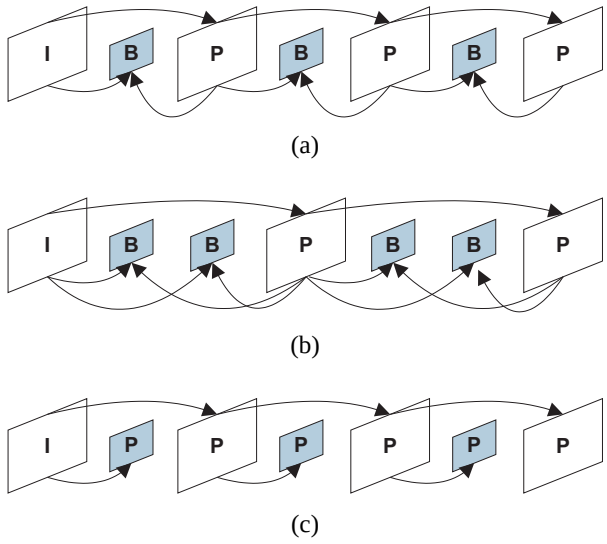


Figure 1. Different uses of NRWZ frames: (a) one B-like NRWZ frame between reference frames; (b) two B-like NRWZ frames; (c) one P-like NRWZ frames.

A. Encoder Architecture

In the NRWZ coding mode, shown in Fig. 2, all the frames in the reference lists and the current frame are decimated by a factor of $2^n \times 2^n$, where n can be selected based on a complexity reduction target. Then, the LR current frame is encoded, generating a reduction in the computational complexity and creating the LR layer bit-stream. The quantization parameter used is the same as that corresponding to the target quality for the frame. Note that the syntax element object list for reference frames is also transformed into an appropriate form for reduced resolution encoding. This operation consists in reducing the resolution of the motion vector field and/or mode decisions for the reference frames. This enables not only Direct-B [1] prediction for B-frames but also fast motion estimation at reduced resolution. In order to create the enhancement layer, i.e. the WZ layer, the encoder computes the difference between the full resolution original frame and the interpolated reconstruction of the LR coded frame, denoted the Laplacian residue. This residual frame is sent to the decoder using a WZ coder. It is easy to see that the encoder complexity for the NRWZ frame is reduced proportionally to the decimation factor (with proper overhead due to decimation, interpolation, and WZ coding operations).

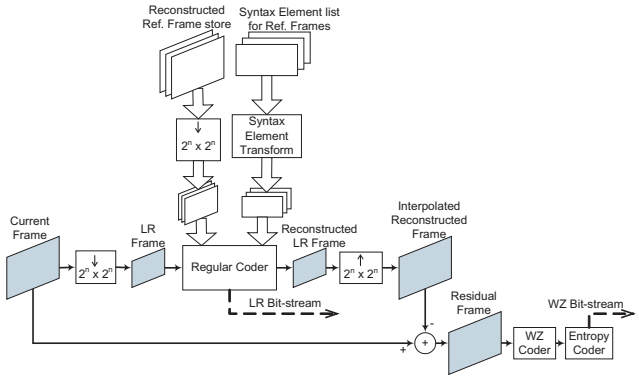


Figure 2. Architecture for the NRWZ coding mode

B. Decoder Architecture

The decoder architecture is presented in Fig. 3. First, the LR image is decoded and interpolated with the same interpolator used at the encoder, generating the decoded base layer. The base layer allows for real-time decoding on power-constrained devices without drop in frame rate, something that is not feasible in traditional temporal Wyner-Ziv coding. The optional process of enhancement begins with the generation of the SI that will be used to decode the WZ information. In the enhancement process, the decoded interpolated frame and the reference frames are used to create what we call the *semi super-resolution* (SSR) version of the current frame. Once the SSR frame is generated, the interpolated LR decoded frame is subtracted from it. The resulting residual frame is the actual SI frame to be used in the WZ decoder. The WZ decoder decodes the WZ bit-stream layer with the SI residual frame acting as a noisy version of the original transmitted residual frame. The decoded residual frame is finally added to the interpolated LR frame to obtain the final decoded frame.

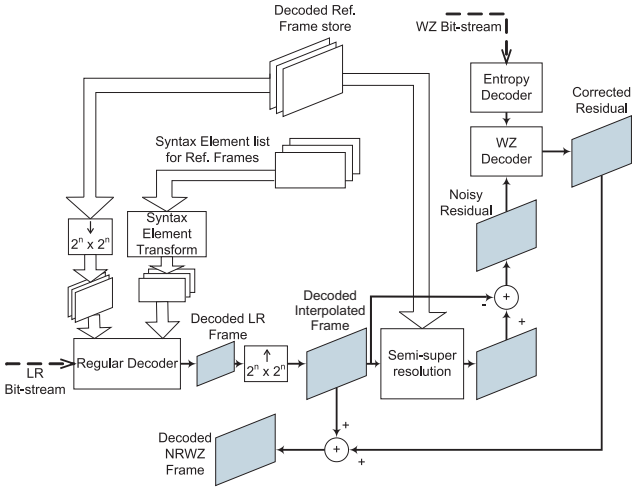


Figure 3. Architecture for the NRWZ decoder

C. Wyner-Ziv Coder

The Wyner-Ziv codec used here works in the block transform domain on the Laplacian residue frame. The model assumes that the block transform coefficients, denoted by random variable X, are Laplacian distributed with standard deviation σ_X. The corresponding coefficient in the SI frame is denoted Y. After the block-transform of the residual frame is computed, the coefficients X are quantized. The random variable that represent the quantized coefficients is denoted by Q, which takes values from the set $\Omega_Q = \{-q_{max}, -q_{max+1}, ..., -1, 0, 1, ..., q_{max-1}, q_{max}\}$. Cosets are next computed on Q with modulus M to obtain the coset random variable C as follows:

$$C = \psi(Q, M) = \begin{cases} Q - M\lfloor Q/M \rfloor, & Q - M\lfloor Q/M \rfloor < M/2 \\ Q - M\lfloor Q/M \rfloor - M, & Q - M\lfloor Q/M \rfloor \geq M/2 \end{cases} \quad (1)$$

The values of C are taken from the set $\Omega_C = \{\lfloor -(M-1)/2 \rfloor, ..., -1, 0, 1, ..., \lfloor (M-1)/2 \rfloor\}$. The QP and M coding parameters are chosen based on an estimate of the noise statistics between the side-information block and the original one, based on a chosen model (see Section IV). In this work, a combination of the number of bits spent to code the corresponding residual block in the LR layer and an edge activity measure in the coded block is used to estimate the model parameters (see Section V).

If the quantization bin q corresponds to interval $[x_l(q), x_h(q)]$, then the probability of the bin $q \in \Omega_Q$, and the probability of a coset index $c \in \Omega_C$ are given by the probability mass functions:

$$p(q) = \int_{x_l(q)}^{x_h(q)} f_X(x)dx \quad (2)$$

$$p(c) = \sum_{q \in \Omega_Q, \psi(Q,M)=c} p(q) = \sum_{q \in \Omega_Q, \psi(Q,M)=c} \int_{x_l(q)}^{x_h(q)} f_X(x)dx, \quad (3)$$

where $f_X(x)$ is the pdf of X. Because the distribution $p(c)$ is symmetric for odd M, has zero as its mode, and decays with increasing magnitude, the entropy coder for Q that already exists in the conventional codec can be reused for C. We note however that a special entropy coder designed specifically for coset indices should be more efficient. In this work we modify and existent entropy coder (CAVLC) in order to improve its performance for the WZ layer.

If Y corresponds to the unquantized SI available only for decoding, then at the decoder, the minimum MSE reconstruction function $\hat{X}_{YC}(y, c)$ based on side information y and received coset index c, is given by

$$\begin{aligned} \hat{X}_{YC}(y, c) &= E(X|Y = y, C = c) \\ &= \frac{\sum_{q \in \Omega_Q, \psi(Q,M)=c} \int_{x_l(q)}^{x_h(q)} x f_{X|Y}(x, y)dx}{\sum_{q \in \Omega_Q, \psi(Q,M)=c} \int_{x_l(q)}^{x_h(q)} f_{X|Y}(x, y)dx}. \end{aligned} \quad (4)$$

In the decoder, the same estimate for the model parameters based on bit rate at the LR layer and edge activity, is obtained. These model parameters not only yield the same QP and M parameters applied during encoding, but are also used to obtain the optimal reconstruction based on (4).

III. Semi Super-Resolution

Our SSR method iteratively computes the super-resolved frame followed by WZ decoding in multiple passes. A block diagram of the process is shown in Fig. 4.

Let the interpolated LR reconstructed frame be F_0. Let $SS(F, FS)$ denote the SSR operation to yield a higher

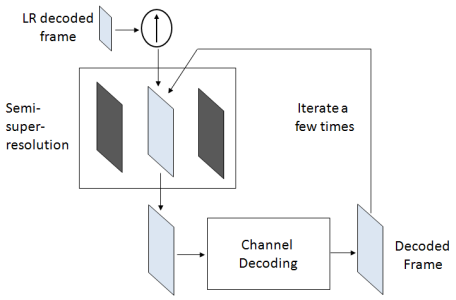


Figure 4. SI generation. Iterate between side-information generation and WZ decoding.

resolution version F^{HR} of F based on the stored frames FS. Also, let $D_{WZ}(RF, b_{WZ})$ denote the WZ decoding operation yielding a corrected version of the residual frame based on our noisy version RF using the WZ layer bit-stream b_{WZ}. Then, iterative decoding comprises the following steps for $i = 0, 1, 2, \ldots, N-1$:

$$F_i^{HR} = SS(F_i, FS) \tag{5}$$

$$F_{i+1} = D_{WZ}(F_i^{HR} - F_0, b_{WZ}) + F_0. \tag{6}$$

In the first iteration, similar to an example-based algorithm [41], we seek to restore the high frequency information of an interpolated block through searching in previously decoded key frames for a similar block, and by adding the high frequency of the chosen block to the interpolated one. The past and future reference frames in the frame-store are low-pass filtered. The low-pass filter is implemented through down-sampling followed by an up-sampling process. The high frequency of the reference frames is the residue between the original frame and its filtered version. If F denotes a frame then $F = L + H$, where L is the decimated and interpolated (filtered) version of F, while H is the residue, or its high frequency.

A block-matching algorithm as described below, is applied on the current decoded frame F_i, for $i = 0, 1, 2, \ldots, N-1$, to obtain the SSR frame F_i^{HR}. For every 8×8 block in F_i, the best sub-pixel motion vectors in the past and future frames are computed to minimize the sum of absolute differences (SAD) between it and the low-pass filtered versions of past and future frames, respectively. Let the best low-pass predictor blocks be denoted as L_p and L_f in the past and future filtered frames respectively. Then, if the SAD of the best predictor computed for a block is lower than a threshold T, the block is updated in the SSR frame F_i^{HR} as follows. Let H_p and H_f denote the high-frequency components of the best predictor blocks. Then in the first iteration, $\alpha^* H_p + (1-\alpha^*)H_f$ is added to the corresponding block in the decoded interpolated frame F_0 to obtain the corresponding block in the super-resolved version F_0^{HR}. In subsequent iterations, the block in the current decoded frame F_i is completely replaced by the unfiltered version of its best predictor: $\alpha^*(H_p + L_p) + (1-\alpha^*)(H_f + L_f)$ to obtain the corresponding block in F_i^{HR}. Further details can be found elsewhere [31].

IV. Choosing Coding Parameters

In order to make an optimal choice of the quantization and modulus parameters $\{QP, M\}$, we assume a general enough statistical model: $Y = \rho X + Z$ [37], [38], where X is a Laplacian distributed transform coefficient with std. dev σ_X, Z is additive Gaussian noise uncorrelated with X with std. dev. σ_Z , and $0 < \rho < 1$ is an attenuation factor expected to decay at higher frequencies. While this is a generalization of the simpler model: $Y = X + Z$, rewriting as $Y/\rho = X + Z/\rho$ the same known estimation procedure [26], [28] can be applied, simply replacing σ_Z^2 with $(\sigma_Z/\rho)^2$ and Y with Y/ρ during decoding. In the rest of this section, we review the optimal parameter choice mechanism for the $Y = X + Z$ model.

A. Memoryless coset codes followed by minimum MSE reconstruction with side-information

The first step is to obtain expressions for expected rate and distortion functions for the memoryless coset codes described in Section II, for a given $\{QP, M\}$ pair. Let R_{YC} be the rate assuming an ideal entropy coder for the coset indices, and let D_{YC} be the distortion given side information y and coset index c. It can be shown [26] that their expected values are given by:

$$E(R_{YC}) = -\sum_{c\in\Omega_C}\left\{\sum_{q\in\Omega_Q:\psi(Q,M)=c}\left[m_X^{(0)}(x_h(q)) - m_X^{(0)}(x_l(q))\right]\right\} \times \log_2\left\{\sum_{q\in\Omega_Q:\psi(Q,M)=c}\left[m_X^{(0)}(x_h(q)) - m_X^{(0)}(x_l(q))\right]\right\}, \tag{7}$$

$$E(D_{YC}) = \sigma_x^2 - \int_{-\infty}^{\infty}\{\sum_{c\in\Omega_C}\frac{\left(\sum_{q\in\Omega_Q:\psi(Q,M)=c}[m_{X|Y}^{(1)}(x_h(q),y) - m_{X|Y}^{(1)}(x_l(q),y)]\right)^2}{\left(\sum_{q\in\Omega_Q:\psi(Q,M)=c}[m_{X|Y}^{(0)}(x_h(q),y) - m_{X|Y}^{(0)}(x_l(q),y)]\right)}\} \times f_Y(y)dy. \tag{8}$$

where we defined $m_X^{(i)}(x) = \int_{-\infty}^{x}\nu^i f_X(\nu)d\nu$ and $m_{X|Y}^{(i)}(x,y) = \int_{-\infty}^{x}\nu^i f_{X|Y}(\nu,y)d\nu$.

A viable coding choice is to just use zero-rate coding, where no information is transmitted (i.e. $QP \to \infty$ or $M = 1$). Then the rate is 0 and it can be shown [26], that the expected distortion based on optimal reconstruction using Y alone is given by:

$$E(D_Y) = \sigma_X^2 - \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} x f_{X|Y}(x,y)dx\right)^2 f_Y(y)dy = \sigma_X^2 - \int_{-\infty}^{\infty} m_{X|Y}^{(1)}(\infty,y)^2 f_Y(y)dy. \tag{9}$$

B. Laplacian source with additive Gaussian noise

Using the model of a Laplacian source, X, with additive Gaussian noise, Z, we have:

$$f_X(x) = \frac{1}{\sqrt{2}\sigma_x}e^{-\left|\frac{x\sqrt{2}}{\sigma_x}\right|}, \tag{10}$$

$$f_Z(z) = \frac{1}{\sqrt{2\pi}\sigma_z}e^{-\frac{1}{2}\left|\frac{z}{\sigma_z}\right|^2}. \tag{11}$$

Defining $\beta(x) = e^{\frac{\sqrt{2}x}{\sigma_x}}$, we have

$$m_X^{(0)}(x) = \begin{cases} \frac{\beta(x)}{2}, x \leq 0 \\ 1 - \frac{1}{2\beta(x)}, x > 0 \end{cases}$$
$$m_X^{(1)}(x) = \begin{cases} \frac{\beta(x)}{2\sqrt{2}}(\sqrt{2}x - \sigma_x), x \leq 0 \\ -\frac{1}{2\sqrt{2}\beta(x)}(\sqrt{2}x + \sigma_x), x > 0 \end{cases} \quad (12)$$

The above moments will be used in the calculation of $E(R_{YC})$ in (7). Further defining:

$$\gamma_1(x) = erf\left(\frac{\sigma_x x - \sqrt{2}\sigma_z^2}{\sqrt{2}\sigma_x\sigma_z}\right),$$
$$\gamma_2(x) = erf\left(\frac{\sigma_x x + \sqrt{2}\sigma_z^2}{\sqrt{2}\sigma_x\sigma_z}\right) \quad (13)$$

and

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt, \quad (14)$$

then, since $Y = X + Z$, we have:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx =$$
$$\frac{1}{2\sqrt{2}\beta(y)\sigma_x} e^{\frac{\sigma_x^2}{\sigma_z^2}}[\gamma_1(y) + 1.0 - \beta(y)^2(\gamma_2(y) - 1.0)], \quad (15)$$

$$f_{X|Y}(x,y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$
$$= \frac{\sqrt{2}\beta(y)}{\sqrt{\pi}\sigma_z} \frac{e^{-\left|\frac{x\sqrt{2}}{\sigma_x}\right| - \frac{1}{2}(\frac{y-x}{\sigma_z})^2 - \frac{\sigma_x^2}{\sigma_z^2}}}{[\gamma_1(y) + 1.0 - \beta(y)^2(\gamma_2(y) - 1.0)]}. \quad (16)$$

Given $f_{X|Y}(x,y)$, the moments $m_{X|Y}^{(i)}(x,y)$ can now be calculated and used for computing $E(D_{YC})$ in (8) by numerical integration. The $erf()$ function can be evaluated based on a $9th$ order polynomial approximation [42]. All the expected rate and distortion values for a memoryless coset code followed by minimum MSE reconstruction can be evaluated based on these results in conjunction with numerical integration of $f_Y(y)$.

C. Optimal parameter choice and model prediction

The above results can be used to compute the RD points for the set of all allowable $\{QP, M\}$ combinations, using (7) and (8). In Fig. 5, we present the RD points so obtained for the specific case of $\sigma_X = 1$ and $\sigma_X = 0.4$. In Fig. 5(a) each RD curve depicted as $'Const\ M\ RD\ Points'$ is generated by fixing M and changing QP at finely sampled intervals of 0.05. The case when $QP \to \infty$ for any M corresponds to the zero-rate case, and yields the RD point $\{0, E(D_Y)\}$ where all the curves start, with $E(D_Y)$ given by (9). Alternatively, this point can also be viewed as the $M = 1$ curve. Figure 5(b) shows exactly the same results using constant QP curves and changing M, starting from 1 upwards. In both Figures, the curve for regular encoding Q using Optimal Reconstruction is depicted as $'EC(Q) + Opt\ recon'$, where EC indicates entropy coded. The curve corresponding to ideal Slepian Wolf coding of the quantization indices (Q) followed by Optimal Reconstruction is also shown ($'SW(Q) + Opt\ recon'$). Note that this curve represents the theoretical upper-bound performance of this framework.

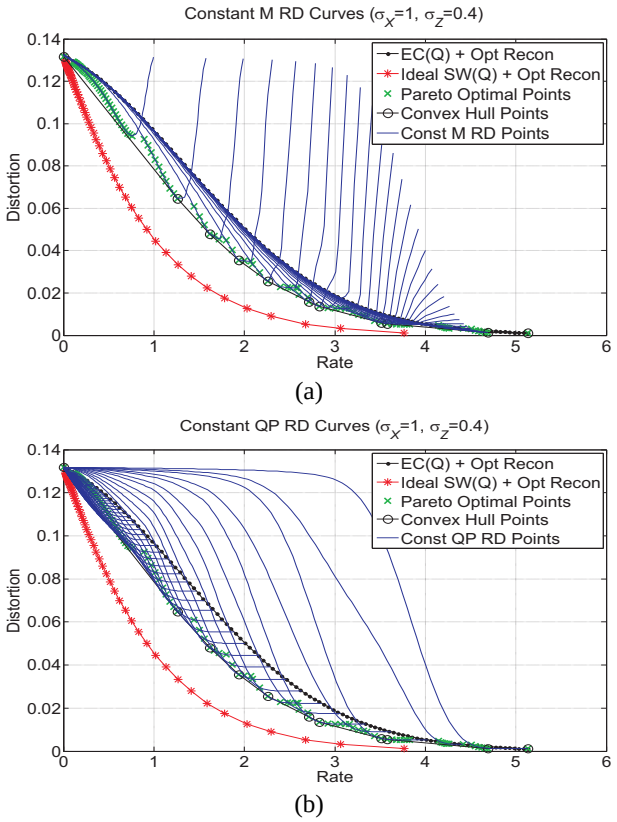


Figure 5. RD curves obtained for $\sigma_X = 1$ and $\sigma_Z = 0.4$, (a) Constant M and varying QP, (b) Constant QP and varying M.

From the Figures, it is obvious that not all choices for QP and M are necessarily good codes, since they may not outperform regular coding. However, as it can also be seen from the Figures, if $\{QP, M\}$ are correctly chosen, coset coding can outperform regular entropy coding of Q with Optimal Reconstruction. The sub-optimal choices for the $\{QP, M\}$ combination can be pruned out by finding the Pareto-optimal set. These points are marked as $'\times'$ in Fig. 5(a) and 5(b). One strategy for parameter choice is to choose the closest Pareto-optimal point to a given target distortion D_t. However, the strategy that yields superior RD performance is to operate on the lower convex hull of the set of points generated by all $\{QP, M\}$ combinations. This set of points, called the convex hull set, is a subset of the points in the Pareto-optimal set and is generally much sparser. These points are marked as $'\circ'$ in Fig. 5(a) and 5(b). In this work we simply chose the closest Pareto-optimal point.

D. Distortion target matching

The goal of the distortion-matched parameter choice process can now be expressed in terms of (7) and (8). If QP_t is the target quantization step-size of regular encoding, used for key frames, we note that it is advantageous, in our framework, to specify the target distortion D_t in terms of QP_t. The expected distortion from regular encoding followed by MSE reconstruction without SI is given by:

$$E(D_Q) = \sigma_X^2 - \sum_{q \in \Omega_Q} \frac{\left(\int_{xl(q)}^{x_h(q)} x f_X(x) dx \right)^2}{\left(\int_{xl(q)}^{x_h(q)} f_X(x) dx \right)}$$
$$= \sigma_X^2 - \sum_{q \in \Omega_Q} \frac{\left(m_X^{(1)}(x_h(q)) - m_X^{(1)}(x_l(q)) \right)^2}{\left(m_X^{(0)}(x_h(q)) - m_X^{(0)}(x_l(q)) \right)} \quad (17)$$

From (17) we obtain D_t for a given QP_t. Thereafter, we search for the optimal code with distortion closest to D_t, but not exceeding it. In practice, this mapping from QP_t to $\{QP, M\}$ can be pre-computed and stored in a normalized table for a given σ_Z and $\sigma_X = 1$, for a range of QP_t values in small incremental steps. Further details about how to generate these tables can be found elsewhere [28].

Reverting back to our $Y = \rho X + Z$ model, the following look-up procedure based on a set of normalized tables is used to obtain the $\{QP^*, M\}$ combination corresponding to a given QP_t^*, if the model parameters $\{\rho, \sigma_X, \sigma_Z\}$ are known. We simply have to (i) evaluate $\sigma_Z/(\rho.\sigma_X)$ to find the normalized look-up table to consult from a set; (ii) find the closest entry in it corresponding to target $QP_t = QP_t^*/\sigma_X$, (iii) read off QP and M; and (iv) scale QP to obtain the final $QP^* = QP \times \sigma_X$. The set of allowable QP values will depend on the conventional codec used (H.264 in our implementation).

V. Correlated Statistic Estimation

In this section, we propose a mechanism to estimate the parameters $\{\rho, \sigma_X, \sigma_Z\}$ in the encoder and decoder for our $Y = \rho X + Z$ model within the proposed spatial scalability framework [37], [38]. The estimated parameters will be used to obtain the coding parameters $\{Qp, M\}$ as described in the previous section.

The model parameters are estimated per frequency band (FB) within a block, where an FB is defined as a diagonal in a transform block. These models are trained based on (X, Y) pair training data collected from a set of training video sequences for each $FB$and QP_t, by running the base layer encoder and SSR processing at the decoder for different values of QP_t in small increments.

A. Estimation of σ_X^2- variance of Laplacian residual coefficients

The variance of a Laplacian residual coefficient (σ_X^2) varies from block to block within a frame. It not only depends on QP_t and FB, but also on the high frequency content within the block. If the original frame has a high edge content it is likely that the error between the decoded interpolated version and the original one would be larger. Even though the exact high frequency content in an original frame is not available at the decoder, an edge activity measure, denoted E, computed on the reconstructed LR block can be used as an indicative parameter to estimate σ_X^2. E is computed as the accumulated sum of the absolute difference between adjacent pixels along lines and columns of a macroblock in the decoded interpolated LR frame. We assume σ_X^2 to be proportional to QP_t^2. Further, analysis of the training data reveals that that it suffices to linearly model the variation of σ_X^2 with E for each FB, so that:

$$\sigma_X^2 = (k_{1,FB} E + k_{2,FB}) QP_t^2 \quad (18)$$

where $k_{i,FB}$ are constants that vary for each frequency band. These parameters are estimated by training based on the training data.

B. Estimation of the correlation parameter

In order to estimate ρ, we use a simplified model assuming that it only depends on QP_t and FB:

$$\rho = f_2(QP_t, FB). \quad (19)$$

If S_{FB,QP_t} represents the training data set for a given QP_t and FB, we estimate:

$$\rho_{FB,QP_t} = argmin_\rho \sum_{(X,Y) \in S_{FB,QP_t}} \left(\|Y - \rho X\|^2 \right) \quad (20)$$

C. Estimation of the variance of the Gaussian noise

We conjecture that σ_Z^2 for a macroblock in the enhancement layer is well indicated by the residual error rate R used to code a co-located 8×8 block in the LR base layer. A higher rate in the LR base layer indicates greater inaccuracy of motion estimation at reduced resolution. Since R depends on QP_t, we can use normalized rate $R_n = R \times QP_t^2$ in order to remove the effect of QP_t. We also assume σ_Z^2 to be proportional to σ_X^2 for a given FB and R_n, and the effect of QP_t and E to be subsumed within σ_X^2. Further, the variation of σ_Z^2 with R_n is linearly modeled for each FB, so that the estimation model is simplified to:

$$\sigma_Z^2 = (k_{3,FB} R_n + k_{4,FB}) \sigma_X^2 \quad (21)$$

Constants $k_{3,FB}$ and $k_{4,FB}$ are estimated by training based on the training data, where for each (X, Y) training pair, the corresponding Z is computed as $Y - \rho X$.

VI. Entropy coder of the WZ layer

This section will describe the modifications done in the context-adaptive variable-length coder (CAVLC) already existent in H.264. The parameters that need to be encoded are the coded block pattern (cbp), that indicates which 8×8 block has non-zero cosets, and the residual data which is coded in 4×4 blocks. The encoding parameters $\{QP, M\}$ are not transmitted, since they are selected at both ends using information from the LR base layer. When it is chosen not to send any cosets for an entire 16×16 macroblock no information is written into the bit-stream, not even the cbp. If, one or more FB are selected for cosets creation ($M \neq 1$) for at least one 4×4 block within a 16×16 macroblock, then the cbp of the macroblock is coded, as well as the residual data. The cbp is encoded like in the conventional CAVLC, using the intra prediction table.

The CAVLC entropy coder module in the H.264/AVC codec is designed to encode zero-runs of the zig-zag ordered residual blocks. The CAVLC takes advantage of several characteristics of the residual quantized blocks. For example, highest frequency non-zero coefficients are often sequences of ± 1, therefore the CAVLC encodes the number of high-frequency ± 1 (trailing ones) in a compact way. CAVLC encodes the residual block in the following order:

1) *the number of non-zero coefficients and trailing ones,*
2) *the sign of each trailing one,*
3) *the levels of the remaining non-zero coefficients,*
4) *total number of zeros before the last coefficient,*
5) *each run of zeros before each non-zero coefficient.*

We will now detail the modifications made in three of the above steps for the coding of the WZ layer. Note that these modifications are used for encoding the AC coefficients. The DC coefficients are encoded as in regular CAVLC, separately.

A. Number of coefficients and trailing ones

The CAVLC encodes both the number of non-zero coefficients and the number of trailing ones as one codeword. CAVLC choses among four look-up tables. Three of them are variable-length code tables and the fourth is a fixed-length code table. The choice of the table depends on the number of non-zero coefficients on the previously coded upper and left blocks. The first variable-length code table is biased towards small numbers, the second towards medium numbers, and the third towards larger numbers. The fixed-length code assigns a 6 bits word to each code. When encoding the cosets we have as additional information the maximum number of non-zero cosets in a 4×4 block. In other words, when $M \neq 1$ during coset creation. This number will be referred as max_C. With the value of max_C, we can modify each of the 3 variable-length code tables. Instead of just 3 different tables, we can create 3 different table groups. Each table group includes code tables with decreasing number of codes entries, hence shorter codes. The number of entries will be limited by max_C.

B. Coefficient Levels

The level of each remaining non-zero coefficient is encoded starting from the highest frequency and going back towards the DC coefficient. In order to encode the levels there are 7 different tables, that are selected based on the last coded level as follows:

1) *Initialize encoding with the first table;*
2) *encode the highest non-zero level;*
3) *if the value of this level is larger than a predefined threshold, use the next table for the next level.*

For the cosets we have the information of the modulus M used to create a particular coset in a FB. A high value of M indicates that it is probably better to use a table for medium or high level values. Therefore, if M is larger than a threshold (10 in our implementation), then encoding starts using the second table instead of the first.

C. Total number of zeros

The total number of all zeros preceding the highest non-zero coefficient is encoded after encoding the levels. There are 15 different code tables used to encode the total number of zeros. Each table depends on the number of non-zero coefficients. Similar to the case of encoding the number of coefficients and trailing ones, when encoding the enhancement layer, we have the information of the maximum number of cosets max_C. This information also indicates the maximum value of the total number of zeros. Then, new tables can be created in addition of the original tables in the CAVLC coder in order to encode the total number of zeros.

VII. Results

The described framework and the proposed SI generation method were implemented on the KTA software for extensions to the state-of-the-art standard H.264/AVC [43]. Results for the H.263+ version can be found elsewhere [27], [28], [32], [38]. For the simulations we used fast motion estimation, with the CAVLC entropy coder for the base layer and D-CAVLC for the WZ layer, without RD optimization (RDO). We also used search range of 16, 2 reference frames, and spatial direct mode type for B-frames. The decimation factor used for the NRWZ frames was 2×2 (half resolution). During the RD tests, the training for the estimation models for $\{\sigma_X, \sigma_Z, \rho\}$ described in Section V, was carried using 15 frames from each of three different sequences (Silent, Mobile and Foreman). The results presented use all 300 frames for sequences: Coastguard, Akiyo, Foreman, Hall monitor, Soccer and Mother-and-Daughter.

In Table I, a comparison of the encoding time between the conventional H.264/AVC and the proposed Wyner-Ziv mode is presented. All the coding tests were made on an Intel Pentium D 915 2.80 GHz Dual Core with Windows OS. The WZ mode was set to operate in $IbPbP$, $IpPpP$ and $IbIbI$ modes where b frames indicate NRWZ B-like frames at half resolution and p is a disposable NRWZ P-like frame, also at half resolution. For conventional H.264/AVC, p_d frames refer to disposable P frames [1]. It can be seen that the mixed resolution $IpPpP$ mode reduces the total encoding time by 37% compared with conventional encoding in Ip_dPp_dP mode. Note that if encoding of an NRWZ p frame at half-resolution is exactly quarter of a disposable p_d frame for conventional H.264/AVC, the complexity reduction would be 37.5%. For the $IbPbP$ mode we achieved a 42% reduction in total encoding time over conventional $IBPBP$. Even lower encoding complexity can be achieved using intra key frames, e.g. in a $IbIbI$ mode, where a reduction of 63% is achieved over conventional $IBIBI$.

Table I
ENCODING TIME COMPARISON FOR COASTGUARD CIF SEQUENCE. (ME: MOTION ESTIMATION)

	H.264-Ip_dPp_dP	*MR-DVC mode-$IpPpP$*
Total time (sec)	259.61	163.41
ME time (sec)	201.12	129.12
	H.264-$IBPBP$	*MR-DVC mode-$IbPbP$*
Total time (sec)	319.61	184.06
ME time (sec)	249.02	142.8
	H.264-$IBIBI$	*MR-DVC mode-$IbIbI$*
Total time (sec)	170.95	63.08
ME time (sec)	132.78	32.44

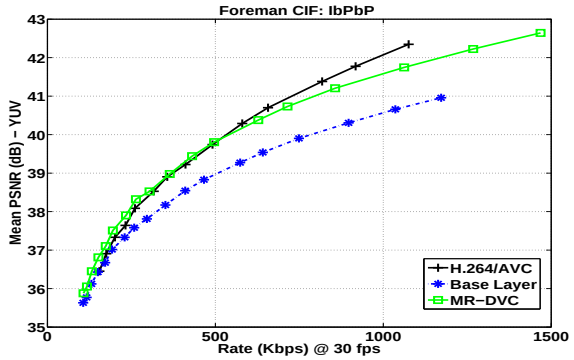
In Table II we show percentage of bit rate reduction in the WZ layer by using the D-CAVLC compare to regular CAVLC

RD results for the $IbPbP$ mode are presented in Fig. 6 for Foreman and Akiyo sequences. Using the mean PSNR as an objective measure [44]. It can be observed that the reversed-complexity WZ mode is competitive. In Fig.7, we compare the performances of our WZ codec in $IpPpP$ mode against a conventional H.264 codec in $IpPpP$ mode for Mother-and-Daughter and Container sequences. It can be observed that, at low rates for many sequences, the SI and especially the WZ modes outperform conventional H.264/AVC in average PSNR, mainly due to accurate SI generation applied to the

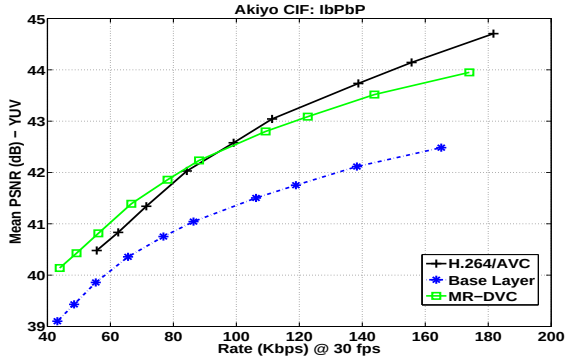
Table II
BIT RATE GAIN FOR THE WZ BIT-STREAM

Sequence	Bitrate Reduction
Foreman	5.91%
Mobile	3.25%
Soccer	2.50%
Coastguard	7.01%
Silent	2.02%
Mother-Daughter	2.43%
Hall Monitor	1.98%
Container	2.05%
Akiyo	1.51%

mixed resolution framework.



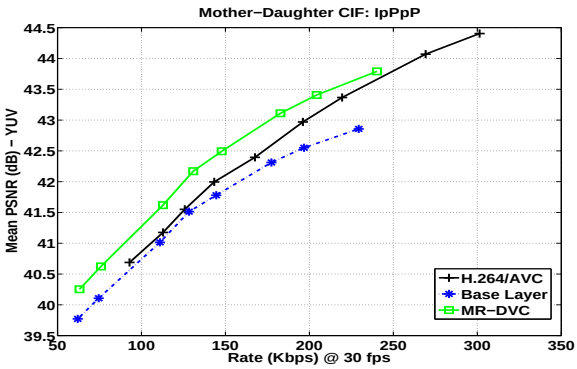
(a)



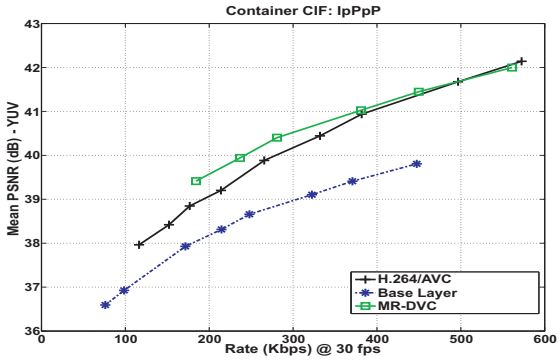
(b)

Figure 6. PSNR results for YUV components at $IBPBP$ mode. Comparing: conventional H.264; the LR base layer after up-sampling; and proposed Mixed Resolution DVC.

For a final comparison, in Fig. 8, we compare our WZ coding mode, working in $IbIbI$ and $IbPbP$ modes, against one of the most popular WZ architectures: the Discover codec [45]. The simulations for the DISCOVER codec were carried with and without RDO, with fast motion estimation, CAVLC entropy coder, and working in $IZIZI$ mode where Z means a WZ frame coded entirely by a Wyner-Ziv coder. It is also important to note that the DISCOVER codec uses sequence dependent QP tables to have a uniform subjective quality through the whole sequence. The DISCOVER currently only works for the luma component. Also, it uses a feedback channel which is not used by the proposed architecture. In terms of complexity, using RDO, the DISCOVER codec takes about 165 seconds to encode the key frames of an entire CIF sequence in average, which makes the encoding complexity



(a)



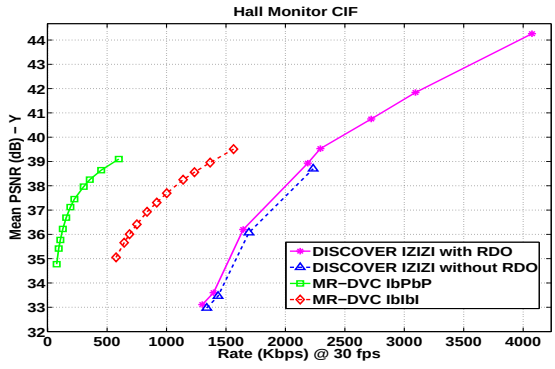
(b)

Figure 7. PSNR results for YUV components at $IpPpP$ mode. Comparing conventional H.264 at $IpPpP$ mode, the low-resolution base layer, and proposed Mixed Resolution DVC.

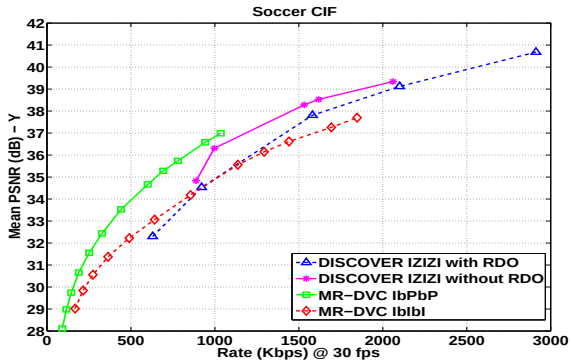
comparable to our scheme without RDO in $IbPbP$ mode. However, without RDO, the DISCOVER codec can encode the key frames in approximately 20 seconds. Our mixed resolution architecture does not achieve such low encoding complexity as indicated in Table I. Nevertheless, as it can be seen in Fig. 8, the proposed architecture achieves better RD performance for the Hall Monitor CIF sequence. For the Soccer CIF sequence DISCOVER outperformed our WZ codec only in $IbIbI$ mode, it is important to note that the DISCOVER codec uses a feedback channel and sequence dependent quantization tables, while our architecture does the rate-allocation automatically.

VIII. CONCLUSION

In this paper, we have presented a mixed resolution Wyner-Ziv video coding framework, and proposed a motion-based semi super-resolution side-information generation mechanism, and a mechanism for selecting coding parameters based on a statistical correlation estimation. Along with a specific entropy coder. The side-information generation and Wyner-Ziv layer decoding are iteratively carried in order to allow for better performance of the Wyner-Ziv decoder. Our Wyner-Ziv coding mode does not use a feedback channel. Hence, a correlation estimation method was proposed in order to choose the right coding parameters. The results show that our mixed resolution framework is competitive and allows reduction of encoding complexity, irrespective of the implementation of the core encoder. It may even outperform regular coding at low bit rates, with lower encoding complexity. Our



(b)



(c)

Figure 8. PSNR results for Y component. Comparing the mixed resolution framework with other WZ architecture.

architecture also enables a low complexity decoding mode if only the base LR layer is decoded.

REFERENCES

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

[2] H. Schwarz, D. Marpe, and T. Wiegand, "SNR-scalable extension of H.264/AVC," *In Proc of IEEE International Conf on Image Processing*, pp. 3113–3116, 2004.

[3] ——, "MCTF and scalability extension of H.264/AVC," *Picture Coding Symposium*, Dezembro 2004.

[4] ——, "Combined scalability support for the scalable extension of H.264/AVC," *In Proc. of IEEE International Conf on Multimedia and Expo.*, Julho 2005.

[5] M. Domanski, L. Blaszak, and S. Mackowiak, "AVC video coders with spatial and temporal scalability," *Picture Coding Symposium*, Abril 2003.

[6] R. Yan, F. Wu, S. Li, and Y. Wang, "Efficient video coding with hybrid spatial and fine-grain SNR scalabilities," *In Proc of SPIE Visual Communications and Image Processing*, vol. 4671, pp. 850–859, 2002.

[7] A.Leontaris and P. Cosman, "Drift-resistant SNR scalable video coding," *IEEE Trans on Image Processing*, vol. 15, no. 8, pp. 2191–2197, Agosto 2006.

[8] G. Cook, J. Nebot, Y.Liu, and E. J. Delp, "Rate-distortion analysis of motion-compensated rate scalable video," *IEEE Trans on Image Processing*, vol. 15, no. 8, pp. 2170–2190, Agosto 2006.

[9] J. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans on Information Theory*, vol. 19, no. 4, pp. 471–480, Julho 1973.

[10] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans on Information Theory*, vol. 2, no. 1, pp. 1–10, Janeiro 1976.

[11] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): design and construction," *in Proc. IEEE Data Compression Conf.*, pp. 158–167, 1999.

[12] A. Aaron, R. Zhang, and B. Girod, "Transform-domain Wyner-Ziv codec for video," *Proc. SPIE Visual Communications and Image Processing*, vol. 5308, pp. 520–528, San Jose, Janeiro 2004.

[13] R. Puri and K. Ramchandran, "PRISM: A new robust video coding architecture based on distributed compression principles," *Allerton Conference on Communications, Control and Computing*, 2002.

[14] Q. Xu and Z. Xiong, "Layered Wyner-Ziv video coding," *IEEE Trans on Image Processing*, vol. 15, no. 12, pp. 3791–3809, Dezembro 2006.

[15] 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*, pp. 1–18, 2006.

[16] M. Tagliasacchi, A. Majumdar, and K. Ramchandran, "A distributed-source-coding based robust spatio-temporal scalable video codec," *Picture Coding Symposium*, San Francisco, Dezembro 2004.

[17] X. Wang and M. T. Orchard, "Desing of trellis codes for source coding with side infotmation at the decoder," *Proceedings of IEEE Data Compression Conference*, pp. 361–370, 2001.

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

[19] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 71–83, Janeiro 2005.

[20] A. M. Aaron, S. Rane, R. Zhang, and B. Girod, "Wyner-Ziv coding for video: applications to compression and error resilience," *Proceedings of the IEEE Data Compression Conference*, pp. 93–102, 2003.

[21] A. Aaron, S. Rane, and B. Girod, "Wyner-ziv video coding with hash-based motion compensation at the receiver," *in Proc. IEEE International Conference on Image Processing*, Outubro 2004.

[22] E. Martinian, A. Vetro, J. Yedidia, J. Ascenso, A. Khisti, and D. Malioutov, "Hybrid distributed video coding using SCA codes," *IEEE Workshop on Multimedia Signal Processing (MMSP)*, pp. 258–261, Outubro 2006.

[23] M. Morbee, J. Prades-Nebot, A. Pizurica, and W. Philips, "Rate allocation algorithm for pixel-domain distributed video coding without feedback channel," *Proceedings of the IEEE ICASSP*, pp. 521–524, Hawaii, 2007.

[24] T. Sheng, G. Hua, H. Guo, J. Zhou, and C. W. Chen, "Rate allocation for transform domain Wyner-Ziv video coding without feedback," *Proceedings of the 16th ACM international conference on Multimedia*, pp. 701–704, 2008.

[25] C. Yaacoub, J. Farah, and B. Pesquet-Popescu, "Feedback channel suppression in distributed video coding with adaptive rate allocation and quantization for multiuser applications," *EUSASIP Journal on Wireless Communications and Networking*, vol. 2008, pp. 1–13, 2008.

[26] B. Macchiavello, D. Mukherjee, and R. L. Queiroz, "Iterative side-information generation in a mixed resolution Wyner-Ziv framework." *to appear in IEEE Trans. on Circuits and Systems for Video Technology*, 2009.

[27] D. Mukherjee, "A robust reversed complexity Wyner-Ziv video codec introducing sign-modulated codes," *HP Labs Technical Report, HPL-2006-80*, Maio 2006.

[28] D. Mukherjee, B. Macchiavello, and R. L. Queiroz, "A simple reversed-complexity Wyner-Ziv video coding mode based on a spatial reduction framework," *Proc. of SPIE Visual Communications and Image Processing*, vol. 6508, pp. 65 081Y1–65 081Y12, Janeiro 2007.

[29] B. Macchiavello, E. Peixoto, and R. L. Queiroz, "A video coding framework with spatial scalability." *Simpsio Brasileiro de Telecomunicaes*, Setembro 2007.

[30] G. Cote, B. Erol, M. Gallant, and F. Kossentini, "H.263+: Video coding at low bit-rates," *IEEE Trans. Circuits Syst. Video Technology*, vol. 8, no. 7, pp. 849–866, Novembro 1998.

[31] B. Macchiavello, F. Brandi, E. Peixoto, R. L. Queiroz, and D. Mukherjee, "Side-information generation for temporally and spatially scalable wyner-ziv codecs," *EURASIP Journal on Image and Video Processing*, vol. 2009, pp. 1–11, 2009.

[32] B. Macchiavello, R. L. Queiroz, and D. Mukherjee, "Motion-based side-inforamtion generation for a scalable Wyner-Ziv video coding," *Proc. of the IEEE International Conference on Imgage Processing*, pp. VI–413–VI–416, 2007.

[33] B. Macchiavello, F. Brandi, R. L. Queiroz, and D. Mukherjee, "Super-resolution applied to distributed video coding with spatial scalability," *Anais do Simpsio Brasileiro de Telecomunicaes*, Rio de Janeiro, 2008.

[34] X. Artigas and L. Torres, "Iterative generation of motion-compensated side information for distributed video coding," *in Proc. IEEE International Conference on Image Processing*, vol. 1, pp. I–833–6, 2005.

[35] J. Ascenso, C. Brites, and F. Pereira, "Motion compensated refinement for low complexity pixel based distributed video coding," *in Proc. of IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 593–598, Septembro 2005.

[36] W. Weerakkody, W. Fernando, J. Martinez, P. Cuenca, and F. Quiles, "An iterative refinement technique for side information generation in DVC," *in Proc. of IEEE International Conference on Multimedia and Expo*, pp. 164–167, Julho 2007.

[37] B. Macchiavello, R. de Queiroz, and D. Mukherjee, "Parameter estimation for an h.264 based distributed video coder," *Proc. of the IEEE International Conference on Image Processing*, pp. 1124–1127, San Diego, 2008.

[38] B. Macchiavello, D. Mukherjee, and R. de Queiroz, "A statistical model for a mixed resolution Wyner-Ziv framework," *Picture Coding Symposium*, Lisboa, 2007.

[39] D. Mukherjee, "Parameter selection for Wyner-Ziv coding of laplacian sources," *Proc. of SPIE Visual Communications and Image Processing*, San Jose, 2007.

[40] 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, 2004.

[41] W. Freeman, T. Jones, and E. Pasztor, "Example-based super-resolution," *IEEE Computer Graphics and Applications*, vol. 22, pp. 56–65, 2002.

[42] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C*, 2nd ed. Cambridge University Press, 1992.

[43] KTA software and JM H.264/AVC reference software: http://iphome.hhi.de/suehring/tml/.

[44] T. Tan, G. Sullivan, and T. Wedi, "Recommended simulation common conditions for coding efficiency experiments," *ITU-T Video Coding Experts Group ITU-T SG16 Q.16 Document VCEG-AE010, 31st Meeting*, Marocco, Jan 2007.

[45] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret, "The discover codec: Architecture, techniques and evaluation," *Picture Coding Symposium*, Novembro 2007.