Hindawi Publishing Corporation
EURASIP Journal on Applied Signal Processing
Volume 2006, Article ID 46747, Pages 1–10
DOI 10.1155/ASP/2006/46747

Distributed Coding of Highly Correlated Image Sequences with Motion-Compensated Temporal Wavelets

Markus Flierl[1] and Pierre Vandergheynst[2]

[1] *Max Planck Center for Visual Computing and Communication, Stanford University, Stanford, CA 94305, USA*
[2] *Signal Processing Institute, Swiss Federal Institute of Technology Lausanne, 1015 Lausanne, Switzerland*

Received 21 March 2005; Revised 27 September 2005; Accepted 4 October 2005

This paper discusses robust coding of visual content for a distributed multimedia system. The system encodes independently two correlated video signals and reconstructs them jointly at a central decoder. The video signals are captured from a dynamic scene, where each signal is robustly coded by a motion-compensated Haar wavelet. The efficiency of the decoder is improved by a disparity analysis of the first image pair of the sequences, followed by disparity compensation of the remaining images of one sequence. We investigate how this scene analysis at the decoder can improve the coding efficiency. At the decoder, one video signal is used as side information to decode efficiently the second video signal. Additional bitrate savings can be obtained with disparity compensation at the decoder. Further, we address the theoretical problem of distributed coding of video signals in the presence of correlated video side information. We utilize a motion-compensated spatiotemporal transform to decorrelate each video signal. For certain assumptions, the optimal motion-compensated spatiotemporal transform for video coding with video side information at high rates is derived. It is shown that the motion-compensated Haar wavelet belongs to this class of transforms. Given the correlation of the video side information, the theoretical bitrate reduction for the distributed coding scheme is investigated. Interestingly, the efficiency of multiview side information is dependent on the level of temporal decorrelation: for a given correlation SNR of the side information, bitrate savings due to side information are decreasing with improved temporal decorrelation.

1. INTRODUCTION

Robust coding of visual content is not just a necessity for multimedia systems with heterogeneous networks and diverse user capabilities. It is also the key for video systems that utilize distributed compression. Let us consider the problem of distributed coding of multiview image sequences. In such a scenario, a dynamic scene is captured by several spatially distributed video cameras and reconstructed at a single central decoder. Ideally, each encoder associated with a camera operates independently and transmits robustly its content to the central decoder. But as each encoder has a priori no specific information about its potential contribution to the reconstruction of the dynamic scene at the central decoder, a highly flexible representation of the visual content is required. In this work, we use a motion-compensated lifted wavelet transform to generate highly scalable bitstreams that can be processed in a coordinated fashion by the central decoder. Moreover, the central decoder receives images of the scene from different viewpoints and is able to perform an analysis of the scene. This analysis helps the central receiver to decode more reliably the incoming robust bitstreams. That is, the decoder is able of content-aware decoding which improves the coding efficiency of the distributed multimedia system that we discuss in the following.

Our distributed system captures a dynamic scene with spatially distributed video cameras and reconstructs it at a single central decoder. Scene information that is acquired by more than one camera can be coded efficiently if the correlation among camera signals is exploited. In one possible compression scenario, encoders of the sensor signals are connected and compress the camera signals jointly. In an alternative compression scenario, each encoder operates independently but relies on a joint decoding unit that receives all coded camera signals. This is also known as distributed source coding. A special case of this scenario is source coding with side information. Wyner and Ziv [1] showed that for certain cases, the encoder does not need the side information to which the decoder has access to achieve the rate distortion bound. Practical coding schemes for our application may utilize a combination of both scenarios and may permit a limited communication between the encoders. But both scenarios have in common that they achieve the same rate distortion bound for certain cases.

Each camera of our system [2] is associated with an encoder utilizing a motion-compensated temporal wavelet

transform [3–5]. With that, we are able to exploit the temporal correlation of each image sequence. In addition, such a wavelet transform provides a scalable representation that permits the desired robust coding of video signals. Interview correlation between the camera signals cannot be exploited as signals from neighboring cameras are not directly available at each encoder. This constraint will be handled by distributed source coding principles. Therefore, the subband coefficients of the wavelet transform are represented by syndromes that are suitable for distributed source coding. A constructive practical framework for the problem of compressing correlated distributed sources using syndromes is presented in [6–8]. To increase the robustness of the syndrome representation, we additionally use nested lattice codes [9]. Syndrome-based distributed source coding is a principle, and several techniques can be employed. For example, [8] investigates memoryless and trellis-based coset construction. For binary sources, turbo codes [10] or low-density parity-check (LDPC) codes [11] increase coding efficiency. Improvements are also possible for nonbinary sources [12–14].

A transform-based approach to distributed source coding for multimedia systems seems promising. The work in [15–18] discusses a framework for the distributed compression of vector sources: first, a suitable distributed Karhunen-Loeve transform is applied and, second, each component is handled by standard distributed compression techniques. That is, each encoder applies a suitable local transform to its input and encodes the resulting components separately in a Wyner-Ziv fashion, that is, treating the compressed description of all other encoders as side information available to the decoder. Similar to that framework, Wyner-Ziv quantization and transform coding of noisy sources at high rates is also investigated in [19, 20]. An application to this framework is the transform-based Wyner-Ziv codec for video frames [21]. In the present article, we capture the efficiency of video coding with video side information based on a high rate approximation. For motion-compensated spatiotemporal transform coding of video with video side information, we derive the optimal transform at high rates, the conditional Karhunen-Loeve transform [22, 23]. For our video signal model, we can show that the motion-compensated Haar wavelet is an optimal transform at high rates.

The coding of multiple views of a dynamic scene is just one part of the problem. The other part addresses which viewpoint will be captured by a camera. Therefore, the underlying problem of our application is sampling and coding of the plenoptic function. The plenoptic function was introduced by Adelson and Bergen [24]. It corresponds to the function representing the intensity and chromaticity of the light observed from every position and direction in the 3D space, at every time. The structure of the plenoptic function determines the correlation in the visual information retrieved from the cameras. This correlation can be estimated using geometrical information such as the position of the cameras and some bounds on the location of the objects [25, 26].

In the present work, two cameras observe the dynamic scene from different viewpoints. Knowing the relative camera position, we are able to compensate the disparity of the reference viewpoint given the current viewpoint. With that, we increase the correlation of the intensity values between the disparity-compensated reference viewpoint and the current viewpoint which lowers the transmission bitrate for a given distortion. Obviously, the higher the correlation between the disparity-compensated reference viewpoint and the viewpoint to be encoded, the lower is the transmission bitrate for a given distortion. As the relative camera positions are not known a priori at the decoder, the first image pair of the two viewpoints is analyzed, and disparity values are estimated. Using these disparity estimates, the decoder can exploit more efficiently the robust representation of the Wyner-Ziv video encoder.

As the present article discusses distributed source coding of highly correlated image sequences, we mention related works of applied research on distributed image coding. For example, [27] enhances analog image transmission systems using digital side information, [28] discusses Wyner-Ziv coding of inter-pictures in video sequences, and [29] investigates distributed compression of light field images. In [30], an uplink-friendly multimedia coding paradigm (PRISM) is proposed. The paradigm is based on distributed source coding principles and renders multimedia systems more robust to transmission losses. Also taking advantage of this paradigm, [31] proposes Wyner-Ziv coding of motion pictures.

The article is organized as follows: Section 2 outlines our distributed coding scheme for two viewpoints of a dynamic scene. We discuss the utilized motion-compensated temporal transform, the cosetencoding of transform coefficients with nested lattice codes, decoding with side information, and enhancing the side information by disparity compensation. Section 3 studies the efficiency of video coding with video side information. Based on a model for transform coded video signals, we address the rate distortion problem with video side information and determine the conditional Karhunen-Loeve transform to obtain performance bounds. The theoretical study finds a tradeoff between the level of temporal decorrelation and the efficiency of decoding with side information. Section 4 provides experimental rate distortion results for decoding of video signals with side information. Moreover, it discusses the relation between the level of temporal decorrelation and the efficiency of decoding with side information.

2. DISTRIBUTED CODING SCHEME

We start with an outline of our distributed coding scheme for two viewpoints of a dynamic scene. We utilize an asymmetric coding scheme; that is, the first viewpoint signal is coded with conventional source coding principles, that is, side information cannot improve decoding of the first viewpoint; and the second viewpoint signal is coded with distributed source coding principles, that is, side information improves decoding of the second viewpoint. The first viewpoint signal is used as video side information to improve decoding of the second viewpoint signal.

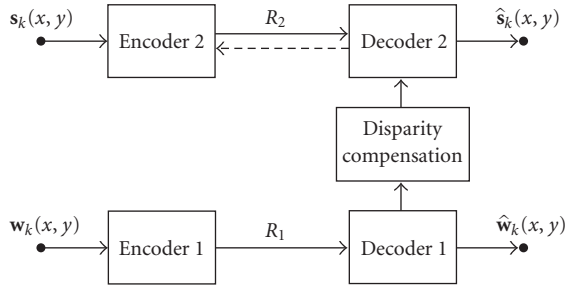

Figure 1: Distributed coding scheme for two viewpoints of a dynamic scene with disparity compensation. The first viewpoint signal is coded at bitrate R_1, the second viewpoint signal at the Wyner-Ziv bitrate R_2.

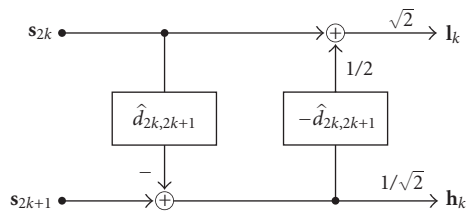

Figure 2: Haar wavelet with motion-compensated lifting steps.

Figure 1 depicts the distributed coding scheme for two viewpoints of a dynamic scene. The dynamic scene is represented by the image sequences $\mathbf{s}_k[x, y]$ and $\mathbf{w}_k[x, y]$. The coding scheme comprises of *Encoder* 1 and *Encoder* 2 that operate independently as well as of *Decoder* 2 that is dependent on *Decoder* 1. The side information for *Decoder* 2 can be improved by considering the spatial camera positions and by performing disparity compensation. As the video signals are not stationary, *Decoder* 2 is decoding with feedback.

2.1. Motion-compensated temporal transform

Each encoder in Figure 1 exploits the correlation between successive pictures by employing a motion-compensated temporal transform for groups of K pictures (GOP). We perform a dyadic decomposition with a motion-compensated Haar wavelet as depicted in Figure 2. The temporal transform provides K output pictures that are decomposed by a spatial 8×8 DCT. The motion information that is required for the motion-compensated wavelet transform is estimated in each decomposition level depending on the results of the lower level. The correlation of motion information between two image sequences is not exploited yet, that is, coded motion vectors are not part of the side information. Figure 2 shows the Haar wavelet with motion-compensated lifting steps. The even frames of the video sequence $\mathbf{s}_{2k}$ are used to predict the odd frames $\mathbf{s}_{2k+1}$ with the estimated motion vector $\hat{d}_{2k,2k+1}$. The prediction step is followed by an update step which uses the negative motion vector as an approximation. We use a block size of 16×16 and half-pel accurate motion, compensation with bilinear interpolation in the prediction step, and select the motion vectors such that they minimize a Lagrangian cost function based on the squared error in the high-band $\mathbf{h}_k$ [5]. Additional scaling factors in low- and high-band are necessary to normalize the transform.

Encoder 1 in Figure 1 encodes the side information for *Decoder* 2 and does not employ distributed source coding principles yet. A scalar quantizer is used to represent the DCT coefficients of all temporal bands. The quantized coefficients are simply run-level encoded. On the other hand, *Encoder* 2 is designed for distributed source coding and uses nested lattice codes to represent the DCT coefficients of all temporal bands.

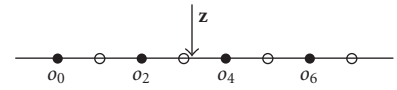

Figure 3: Coset coding of transform coefficients, where *Encoder* 2 transmits at a rate R_{TX} of 1 bit per transform coefficient.

2.2. Nested lattice codes for transform coefficients

The 8×8 DCT coefficients of *Encoder* 2 are represented by a 1-dimensional nested lattice code [9]. Further, we construct cosets in a memoryless fashion [8].

Figure 3 explains the coset-coding principle. Assume that *Encoder* 2 transmits at a rate R_{TX} of 1 bit per transform coefficient, and utilizes two cosets $\mathcal{C}_{1,0} = \{o_0, o_2, o_4, o_6\}$ and $\mathcal{C}_{1,1} = \{o_1, o_3, o_5, o_7\}$ for encoding. Now, the transform coefficient o_4 will be encoded and the encoder sends one bit to signal coset $\mathcal{C}_{1,0}$. With the help of the side information coefficient $\mathbf{z}$, the decoder is able to decode o_4 correctly. If *Encoder* 2 does not send any bit, the decoder will decode o_3 and we observe a decoding error.

Consider the 64 transform coefficients $\mathbf{c}_i$ of the 8×8 DCT at *Encoder* 2. The correlation between the ith transform coefficient $\mathbf{c}_i$ at *Encoder* 2 and the ith transform coefficient of the side information $\mathbf{z}_i$ depends strongly on the coefficient index i. In general, the correlation between corresponding DC coefficients ($i = 0$) is very high, whereas the correlation between corresponding high-frequency coefficients decreases rapidly. To encounter the problem of varying correlation, we adapt the transmission rate R_{TX} to each transform coefficient. For weakly correlated coefficients, a higher transmission rate has to be chosen.

Adapting the transmission rate to the actual correlation is accomplished with nested lattice codes [9]. The idea of nested lattices is, roughly, to generate diluted versions of the original coset code. As we use uniform scalar quantization, we consider the 1-dimensional lattice. Figure 4 depicts the fine code $\mathcal{C}_0$ in the Euclidean space with minimum distance Q. $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$ are nested codes with the νth coset $\mathcal{C}_{\mu,\nu}$ of $\mathcal{C}_\mu$ relative to $\mathcal{C}_0$. The nested codes are coarser and the union of their cosets gives the fine code $\mathcal{C}_0$, that is, $\bigcup_\nu \mathcal{C}_{1,\nu} = \mathcal{C}_0$.

The binary representation of the quantized transform coefficients determines its coset representation in the nested lattice. If the transmission rate for a coefficient is $R_{\mathrm{TX}} = \mu$,

then the μ least significant bits of the binary representation determine the νth coset $\mathcal{C}_{\mu,\nu}$. For highly correlated coefficients, the number of required cosets and, hence, the transmission rate are small. To achieve efficient entropy coding of the binary representation of all 64 transform coefficients, we define bitplanes. Each bitplane is run-length encoded and transmitted to *Decoder* 2 upon request.

2.3. Decoding with side information

At *Encoder* 2, the quantized transform coefficients are represented with 10 bitplanes, where 9 are used for encoding the absolute value, and one is used for the sign. *Encoder* 2 is able to provide the full bitplanes, independent of any side information at the *Decoder* 2. *Encoder* 2 is also able to receive a bitplane mask to weight the current bitplane. The masked bitplane is run-length encoded and transmitted to *Decoder* 2.

Given the side information at *Decoder* 2, masked bitplanes are requested from *Encoder* 2. For that, *Decoder* 2 sets the bitplane mask to indicate the bits that are required from *Encoder* 2. Dependent on the received bitplane mask, *Encoder* 2 transmits the weighted bitplane utilizing run-length encoding. *Decoder* 2 attempts to decode the already received bitplanes with the given side information. In case of decoding error, *Decoder* 2 generates a new bitplane mask and requests a further weighted bitplane.

Decoder 2 has the following options for each mask bit: if a bit in the bitplane is not needed, the mask value is 0. The mask value is 1 if the bit is required for error-free decoding. If the information at the decoder is not sufficient for this decision, the mask is set to 2 and the encoded transform coefficient that is used as side information is transmitted to *Encoder* 2. With this side information $\mathbf{z}_i$ for the ith transform coefficient $\mathbf{c}_i$, *Encoder* 2 is able to determine its best transmission rate $\mu = R_{\mathrm{TX}}[i]$ and coset $\mathcal{C}_{\mu,\nu}$. This information is incorporated into the current bitplane and transmitted to *Decoder* 2: bits that are not needed for error-free decoding are marked with 0. Further, 1 indicates that the bit is needed and its value is 0, and 2 indicates that the bit is needed with value 1.

Decoder 2 aims to estimate the ith transform coefficient $\hat{\mathbf{c}}_i$ based on the current transmission rate $\mu = R_{\mathrm{TX}}[i]$, the partially received coset $\mathcal{C}_{\mu,\nu}$, and the side information $\mathbf{z}_i$:

$$\hat{\mathbf{c}}_i = \underset{\mathbf{c}_i \in \mathcal{C}_{\mu,\nu}}{\operatorname{argmin}} \left[\mathbf{c}_i - \mathbf{z}_i\right]^2 \quad \text{given } \mu = R_{\mathrm{TX}}[i]. \tag{1}$$

With increasing number of received bitplanes, that is, increasing transmission rate $R_{\mathrm{TX}}[i]$, this estimate gets more accurate and stays definitely constant for rates beyond the critical transmission rate $R^*_{\mathrm{TX}}[i]$. Therefore, a simple decoding algorithm is as follows: an additional bit is required if the estimated coefficient changes its value when the transmission rate increases by 1. An unchanged value for an estimated coefficient is just a necessary condition for having achieved the critical transmission rate. This condition is not sufficient for error-free decoding and, in this case, *Encoder* 2 has to determine the critical transmission rate to resolve any ambiguity.

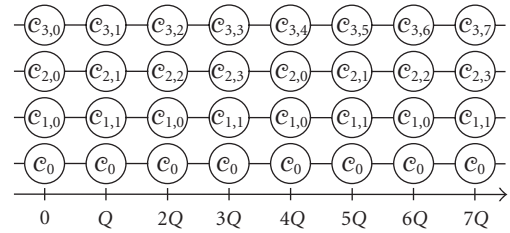

Figure 4: Nested lattices. The 1-dimensional fine code $\mathcal{C}_0$ is embedded into the Euclidean space with minimum distance Q. $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$ are nested codes with the νth coset $\mathcal{C}_{\mu,\nu}$ of $\mathcal{C}_\mu$ relative to $\mathcal{C}_0$.

Note that *Decoder* 2 receives the coded information in bitplane units, starting with the plane of least significant bits. With each new bitplane, *Decoder* 2 utilizes a coarser lattice where the number of cosets as well as the minimum Euclidean distance increases exponentially.

Depending on the quality of the side information, *Decoder* 2 gives feedback to *Encoder* 2 about the status of its decoding attempts. If the correlation of the side information is high, *Decoder* 2 will decode successfully without sending much feedback information. On the other hand, weakly correlated side information will cause decoding errors at *Decoder* 2 and more feedback information is sent to *Encoder* 2 until *Decoder* 2 is successful. That is, inefficient side information is compensated by the feedback.

2.4. Disparity-compensated side information

To improve the efficiency of *Decoder* 2, the side information from *Decoder* 1 is disparately compensated in the image domain. If the camera positions are unknown, the coding system estimates the disparity information from sample frames. During this calibration process, the side information for *Decoder* 2 is less correlated, and *Encoder* 2 has to transmit at a higher bitrate. Our system utilizes block-based estimates of the disparity values which are constant for all corresponding image pairs in the stereoscopic sequence. We estimate the disparity from the first pair of images in the sequences. The right image is subdivided horizontally into 4 segments and vertically into 6 segments. For each of the 24 blocks in the right image, we estimate half-pel accurate disparity vectors. Intensity values for half-pel positions are obtained by bilinear interpolation. The estimated disparity vectors are applied in the image domain and improve the side information in the transform domain. For our experiments, the camera positions are unaltered in time. Therefore, the disparity information is estimated from the first frames of the image sequences and is reused for disparity compensation of the remaining images.

3. EFFICIENCY OF VIDEO CODING WITH SIDE INFORMATION

In this section, we outline a signal model to study video coding with video side information in more detail. We derive

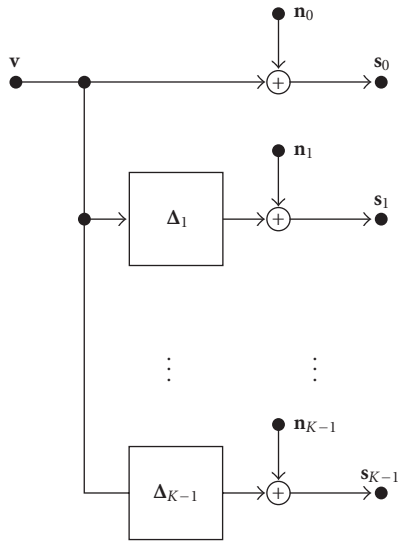

Figure 5: Signal model for a group of K pictures.

performance bounds and compare to coding without video side information.

3.1. Model for transform-coded video signals

We build upon a model for motion-compensated subband coding of video that is outlined in [5, 32]. Let the video pictures $\mathbf{s}_k = \{\mathbf{s}_k[x, y],\ (x, y) \in \Pi\}$ be scalar random fields over a two-dimensional orthogonal grid Π with horizontal and vertical spacing of 1.

As depicted in Figure 5, we assume that the pictures $\mathbf{s}_k$ are shifted versions of the model picture $\mathbf{v}$ and degraded by independent additive white Gaussian noise $\mathbf{n}_k$ [5]. $\boldsymbol{\Delta}_k$ is the displacement error in the kth picture, statistically independent from the model picture $\mathbf{v}$ and the noise $\mathbf{n}_k$ but correlated to other displacement errors. We assume a 2D normal distribution with variance $\sigma_{\boldsymbol{\Delta}}^2$ and zero mean where the x- and y-components are statistically independent. As outlined in [5], it is assumed that the true displacements are known at the encoder. Consequently, the true motion can be set to zero without loss of generality. Therefore, only the displacement error but not the true motion is considered in the model.

From [5], we adopt the matrix of the power spectral densities of the pictures $\mathbf{s}_k$ and normalize it with respect to the power spectral density of the model picture $\mathbf{v}$. We write it also with the identity matrix I and the matrix $\mathbf{1}\mathbf{1}^T$ with all entries equal to 1. Note that ω denotes the 2D frequency:

$$\frac{\Phi_{\mathbf{ss}}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} = \begin{pmatrix} 1+\alpha(\omega) & P(\omega) & \cdots & P(\omega) \\ P(\omega) & 1+\alpha(\omega) & \cdots & P(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ P(\omega) & P(\omega) & \cdots & 1+\alpha(\omega) \end{pmatrix} = \big[1+\alpha(\omega)-P(\omega)\big]I + P(\omega)\mathbf{1}\mathbf{1}^T, \tag{2}$$

and $\alpha = \alpha(\omega)$ is the normalized power spectral density of the noise $\Phi_{\mathbf{n}_k\mathbf{n}_k}(\omega)$ with respect to the model picture $\mathbf{v}$:

$$\alpha(\omega) = \frac{\Phi_{\mathbf{n}_k\mathbf{n}_k}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} \quad \text{for } k = 0, 1, \ldots, K-1. \tag{3}$$

It captures the error of the optimal displacement estimator and will be statistically independent of the model picture. $P = P(\omega)$ is the characteristic function of the continuous 2D Gaussian displacement error. For details, please see (3)–(6) in [5],

$$P(\omega) = E\{e^{-j\omega^T\boldsymbol{\Delta}_k}\} = e^{-(1/2)\omega^T\omega\sigma_{\boldsymbol{\Delta}}^2}. \tag{4}$$

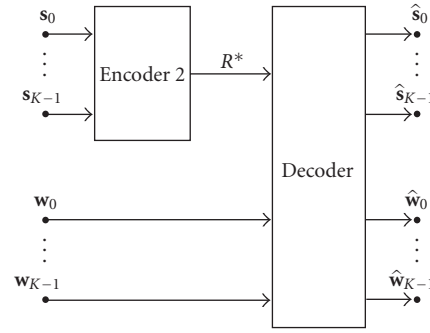

Figure 6: Coding of K pictures $\mathbf{s}_k$ at rate R^* with side information of K pictures $\mathbf{w}_k$ at the decoder.

3.2. Rate distortion with video side information

Now, we consider the video coding scheme in Figure 1 at high rates such that the reconstructed side information approaches the original side information $\widehat{\mathbf{w}}_k \to \mathbf{w}_k$. With that, we have a Wyner-Ziv scheme (Figure 6), and the rate distortion function R^* of *Encoder* 2 is bounded by the conditional rate distortion function [1].

In the following, we assume very accurate optimal disparity compensation and consider only disparity compensation errors. We model the side information as a noisy version of the video signal to be encoded, that is, $\mathbf{w}_k = \mathbf{s}_k + \mathbf{u}_k$, and assume that the noise $\mathbf{u}_k$ is also Gaussian with variance $\sigma_{\mathbf{u}}^2$ and independent of $\mathbf{s}_k$. Further, the side information noise $\mathbf{u}_k$ is assumed to be temporally uncorrelated. This is realistic as the video side information is captured by a second camera which provides temporally successive images that are corrupted by statistically independent camera noise. In this case, the matrix of the power spectral densities of the side information pictures is simply $\Phi_{\mathbf{ww}}(\omega) = \Phi_{\mathbf{ss}}(\omega) + \Phi_{\mathbf{uu}}(\omega)$ with the matrix of the normalized power spectral densities of the side information noise:

$$\frac{\Phi_{\mathbf{uu}}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} = \begin{pmatrix} \gamma(\omega) & 0 & \cdots & 0 \\ 0 & \gamma(\omega) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma(\omega) \end{pmatrix} = \gamma(\omega)I. \tag{5}$$

$\gamma = \gamma(\omega)$ is the normalized power spectral density of the side information noise $\Phi_{\mathbf{u}_k\mathbf{u}_k}(\omega)$ with respect to the model picture $\mathbf{v}$:

$$\gamma(\omega) = \frac{\Phi_{\mathbf{u}_k\mathbf{u}_k}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} \quad \text{for } k = 0, 1, \ldots, K-1. \tag{6}$$

With these assumptions, the rate distortion function R^* of *Encoder* 2 is equal to the conditional rate distortion function [1]. Now, it is sufficient to use the conditional Karhunen-Loeve transform to code video signals with side information and achieve the conditional rate distortion function.

3.3. Conditional Karhunen-Loeve transform

In the case of motion-compensated transform coding of video with side information, the conditional Karhunen-Loeve transform is required to obtain the performance bounds. We determine the well-known conditional power spectral density matrix $\Phi_{\mathbf{s}|\mathbf{w}}(\omega)$ of the video signal $\mathbf{s}_k$ given the video side information $\mathbf{w}_k$:

$$\Phi_{\mathbf{s}|\mathbf{w}}(\omega) = \Phi_{\mathbf{ss}}(\omega) - \Phi^H_{\mathbf{ws}}(\omega)\Phi^{-1}_{\mathbf{ww}}(\omega)\Phi_{\mathbf{ws}}(\omega). \tag{7}$$

With the model in Section 3.1, the assumptions in Section 3.2, and the mathematical tools presented in [33], we obtain for the normalized conditional spectral density matrix

$$\begin{aligned}\frac{\Phi_{\mathbf{s}|\mathbf{w}}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} &= \frac{A(\omega)}{A(\omega)+\gamma(\omega)}\gamma(\omega)I + \frac{P(\omega)}{A(\omega)+\gamma(\omega)} \\ &\quad \cdot \frac{\gamma(\omega)}{A(\omega)+KP(\omega)+\gamma(\omega)}\gamma(\omega)\mathbf{1}\mathbf{1}^T,\end{aligned} \tag{8}$$

where $A(\omega) = 1 + \alpha(\omega) - P(\omega)$. For our signal model, the conditional Karhunen-Loeve transform is as follows: the first eigenvector just adds all components and scales with $1/\sqrt{K}$. For the remaining eigenvectors, any orthonormal basis can be used that is orthogonal to the first eigenvector. The Haar wavelet that we use for our coding scheme meets these requirements. Finally, K eigendensities are needed to determine the performance bounds:

$$\begin{aligned}\frac{\Lambda^*_0(\omega)}{\Phi_{\mathbf{vv}}(\omega)} &= \frac{A(\omega)+KP(\omega)\gamma(\omega)/\big(A(\omega)+KP(\omega)+\gamma(\omega)\big)}{A(\omega)+\gamma(\omega)}\gamma(\omega), \\ \frac{\Lambda^*_k(\omega)}{\Phi_{\mathbf{vv}}(\omega)} &= \frac{A(\omega)}{A(\omega)+\gamma(\omega)}\gamma(\omega), \quad k = 1, 2, \ldots, K-1.\end{aligned} \tag{9}$$

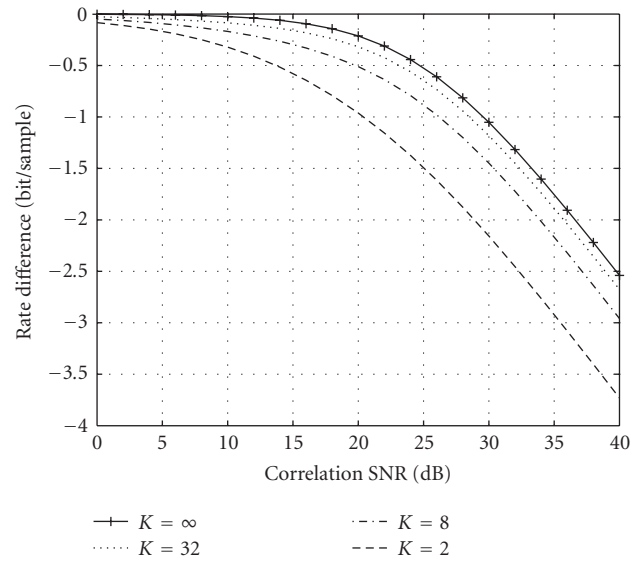

Figure 7: Rate difference between motion-compensated transform coding with side information and without side information versus correlation SNR for groups of K pictures. The displacement inaccuracy β is -1 (half-pel accuracy) and the residual noise is -30 dB.

3.4. Coding gain due to side information

With the conditional eigendensities, we are able to determine the coding gain due to side information. We normalize the conditional eigendensities $\Lambda^*_k(\omega)$ with respect to the eigendensities $\Lambda_k(\omega)$ that we obtain for coding without side information as $\Lambda^*_k(\omega) \to \Lambda_k(\omega)$ for $\gamma(\omega) \to \infty$:

$$\begin{aligned}\frac{\Lambda^*_0(\omega)}{\Lambda_0(\omega)} &= \frac{\gamma(\omega)}{A(\omega)+\gamma(\omega)} \\ &\quad \cdot \frac{A(\omega)+KP(\omega)\gamma(\omega)/\big(A(\omega)+KP(\omega)+\gamma(\omega)\big)}{A(\omega)+KP(\omega)}, \\ \frac{\Lambda^*_k(\omega)}{\Lambda_k(\omega)} &= \frac{\gamma(\omega)}{A(\omega)+\gamma(\omega)}, \quad k = 1, 2, \ldots, K-1.\end{aligned} \tag{10}$$

The rate difference is used to measure the improved compression efficiency for each picture k in the presence of side information:

$$\Delta R^*_k = \frac{1}{4\pi^2}\iint_{-\pi}^{\pi}\frac{1}{2}\log_2\left(\frac{\Lambda^*_k(\omega)}{\Lambda_k(\omega)}\right)d\omega. \tag{11}$$

It represents the maximum bitrate reduction (in bit/sample) possible by optimum encoding of the eigensignal with side information, compared to optimum encoding of the eigensignal without side information for Gaussian wide-sense stationary signals for the same mean-square reconstruction error. The overall rate difference ΔR^* is the average over all K eigensignals [32, 34].

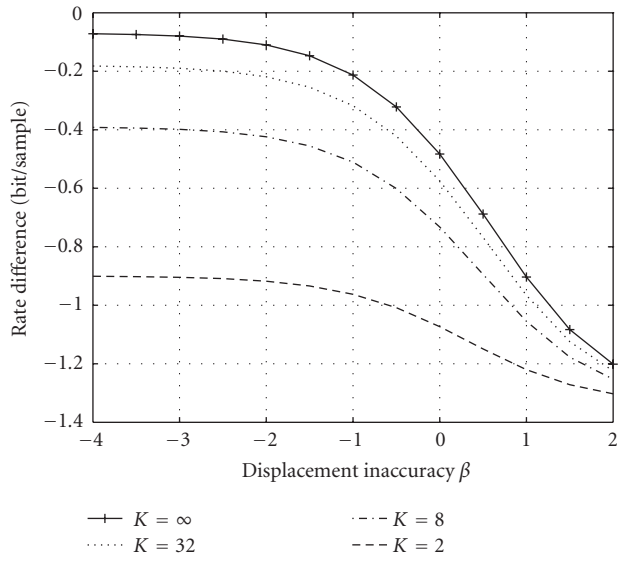

Figure 8: Rate difference between motion-compensated transform coding with side information and without side information versus displacement inaccuracy β for groups of K pictures. The residual noise is -30 dB and the correlation-SNR is 20 dB.

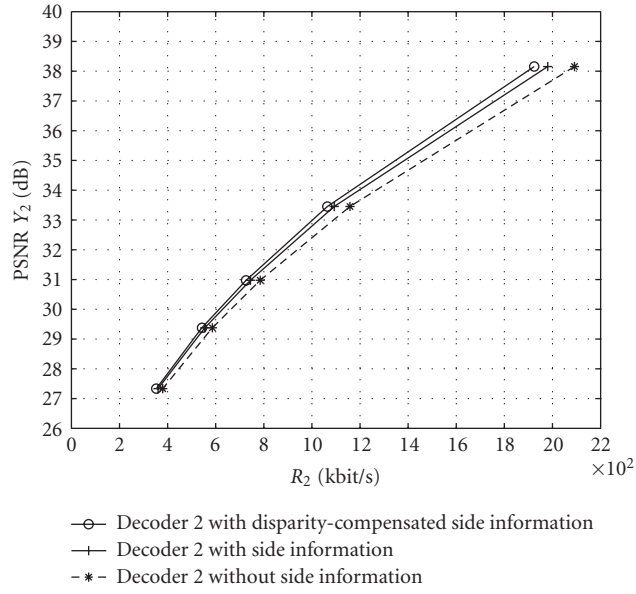

Figure 9: Luminance PSNR versus total bitrate at *Decoder* 2 for the sequence *Funfair* 2 (right view). Compared are decoding with disparity-compensated side information, decoding with coefficient side information only, and decoding without side information. For all cases, groups of $K = 32$ pictures are used.

Figure 7 depicts the overall rate difference for a residual noise level RNL $= 10\log_{10}(\sigma_{\mathbf{n}}^2)$ of -30 dB over the c-SNR $= 10\log_{10}([\sigma_{\mathbf{v}}^2 + \sigma_{\mathbf{n}}^2]/\sigma_{\mathbf{u}}^2)$ for a displacement inaccuracy $\beta = \log_2(\sqrt{12}\sigma_{\Delta}) = -1$. Note that the variance of the model picture $\mathbf{v}$ is normalized to $\sigma_{\mathbf{v}}^2 = 1$. We observe for a given correlation SNR of the side information that larger bitrate savings are achievable if the GOP size K is smaller. The experimental results in Figures 10 and 12 will verify this observation. Finally, for highly correlated video signals, the gain due to side information increases by 1 bit/sample if the c-SNR increases by 6 dB.

Figure 8 depicts the overall rate difference for a residual noise level RNL $= 10\log_{10}(\sigma_{\mathbf{n}}^2)$ of -30 dB over the displacement inaccuracy $\beta = \log_2(\sqrt{12}\sigma_{\Delta})$ for a c-SNR $= 10\log_{10}([\sigma_{\mathbf{v}}^2 + \sigma_{\mathbf{n}}^2]/\sigma_{\mathbf{u}}^2)$ of 20 dB. Again, the variance of the model picture $\mathbf{v}$ is normalized to $\sigma_{\mathbf{v}}^2 = 1$. We observe that for $K = 32$, half-pel accurate motion compensation ($\beta = -1$), and a c-SNR of 20 dB, the rate difference is limited to -0.3 bit/sample. Also, the bitrate savings due to side information increase for less accurate motion compensation. That is, there is a tradeoff between the gain due to accurate motion compensation and side information. Practically speaking, less accurate motion compensation reduces the coding efficiency of the encoder, and with that, its computational complexity, but improved side information may compensate for similar overall efficiency.

4. EXPERIMENTAL RESULTS

For the experiments, we select the stereoscopic MPEG-4 sequences *Funfair* and *Tunnel* in QCIF resolution. We divide each view with 224 frames at 30 fps into groups of $K = 32$ pictures. The GOPs of the left view are encoded with *Encoder* 1 at high quality by setting the quantization parameter $QP = 2$, where $Q = 2QP$. This coded version of the left view is used for disparity compensation. The compensated frames provide the side information for *Decoder* 2 to decode the right view.

Figures 9 and 11 show the luminance PSNR over the total bitrate of the distributed codec *Encoder* 2 for the sequences *Funfair* 2 and *Tunnel* 2, respectively. The sequences are the right views of the stereoscopic sequences. The rate distortion points are obtained by varying the quantization parameter for the nested lattice in *Encoder* 2. When compared to decoding without side information, decoding with coefficient side information reduces the bitrate of *Funfair* 2 by up to 5% and that of *Tunnel* 2 by up to 8%. Decoding with disparity-compensated side information reduces the bitrate of *Funfair* 2 by up to 8%. The block-based disparity compensation has limited accuracy and is not beneficial for *Tunnel* 2. But utilizing more accurate geometrical information about the scene will improve the side information for *Decoder* 2 and, hence, will further reduce the bitrate of *Encoder* 2.

Figures 10 and 12 show the bitrate difference between decoding with side information and decoding without side information over the luminance PSNR at *Decoder* 2 for the sequences *Funfair* 2 (right view) and *Tunnel* 2 (right view), respectively. The bitrate savings due to side information are depicted for weak temporal filtering with $K = 8$ pictures per GOP and strong temporal filtering with $K = 32$ pictures per GOP. Note that both the coded signal (right view) and the side information (left view) are encoded with the same GOP length K. It is observed that strong temporal filtering results in lower bitrate savings due to side information when

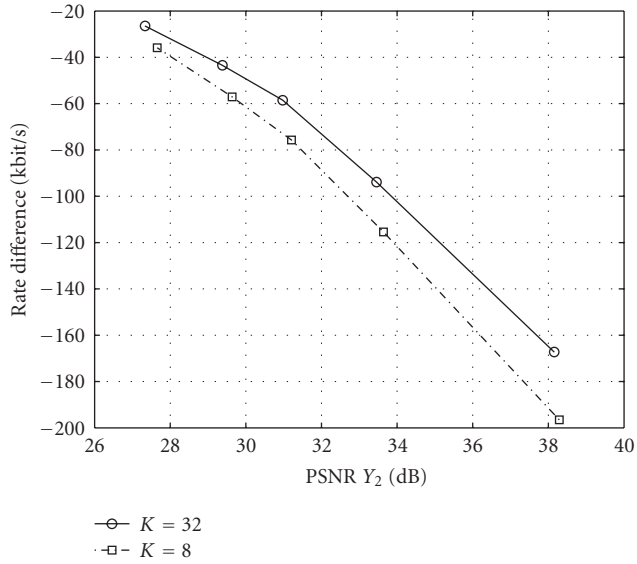

Figure 10: Bitrate difference versus luminance PSNR at *Decoder* 2 for the sequence *Funfair* 2 (right view). The rate difference is the bitrate for decoding with side information minus the bitrate for decoding without side information and reflects the bitrate savings due to decoding with side information. Smaller bitrate savings are observed for strong temporal decorrelation ($K = 32$) when compared to the bitrate savings for weak temporal decorrelation ($K = 8$).

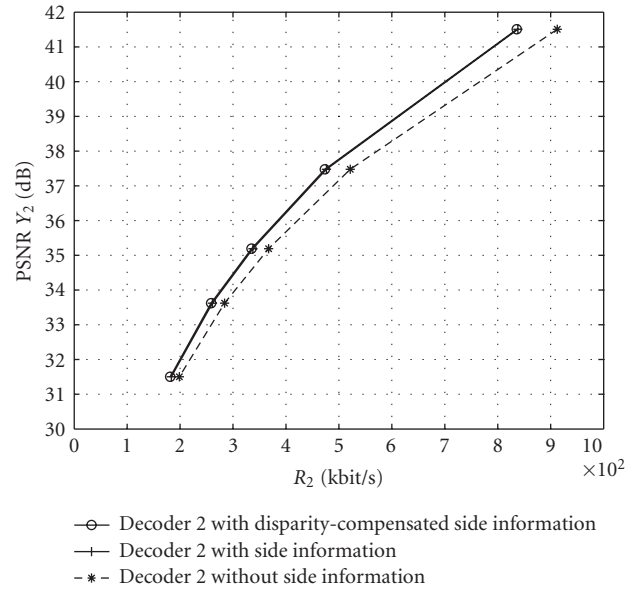

Figure 11: Luminance PSNR versus total bitrate at *Decoder* 2 for the sequence *Tunnel* 2 (right view). Compared are decoding with disparity-compensated side information, decoding with coefficient side information only, and decoding without side information. For all cases, groups of $K = 32$ pictures are used.

compared to the bitrate savings due to side information for weaker temporal filtering. Obviously, there is a tradeoff between the level of temporal decorrelation and the efficiency

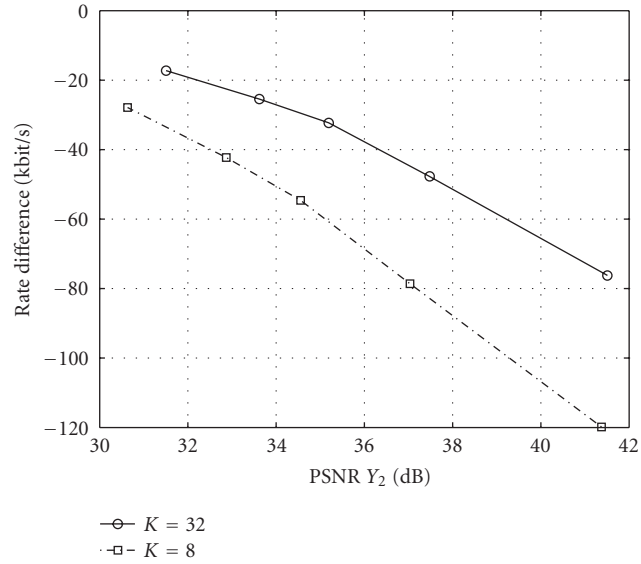

Figure 12: Bitrate difference versus luminance PSNR at *Decoder* 2 for the sequence *Tunnel* 2 (right view). The rate difference is the bitrate for decoding with side information minus the bitrate for decoding without side information and reflects the bitrate savings due to decoding with side information. Smaller bitrate savings are observed for strong temporal decorrelation ($K = 32$) when compared to the bitrate savings for weak temporal decorrelation ($K = 8$).

of multiview side information. This tradeoff is also found in the theoretical investigation on the efficiency of video coding with side information.

5. CONCLUSIONS

This paper discusses robust coding of visual content for a distributed multimedia system. The distributed system compresses two correlated video signals. The coding scheme is based on motion-compensated temporal wavelets and transform coding of temporal subbands. The scalar transform coefficients are represented by a nested lattice code. For this representation, we define bitplanes and encode these with run-length coding. As the correlation of the transform coefficients is not stationary, we decode with feedback and adapt the coarseness of the code to the actual correlation. Also, we investigate how scene analysis at the decoder can improve the coding efficiency of the distributed system. We estimate the disparity between the two views and perform disparity compensation. With disparity-compensated side information, we observe up to 8% bitrate savings over decoding without side information.

Finally, we investigate theoretically motion-compensated spatiotemporal transforms. We derive the optimal motion-compensated spatiotemporal transform for video coding with video side information at high rates. For our video signal model, we show that the motion-compensated Haar wavelet is an optimal transform at high rates. Given the correlation of the video side information, we also investigate

the theoretical bitrate reduction for the distributed coding scheme. We observe a tradeoff in coding efficiency between the level of temporal decorrelation and the efficiency of multiview side information. A similar tradeoff is found between the level of accurate motion compensation and the efficiency of multiview side information.

ACKNOWLEDGMENT

This work has been supported, in part, by the Max Planck Center for Visual Computing and Communication.

REFERENCES

[1] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, 1976.

[2] M. Flierl and P. Vandergheynst, "Distributed coding of dynamic scenes with motion-compensated wavelets," in *Proceedings of IEEE 6th Workshop on Multimedia Signal Processing (MMSP '04)*, pp. 315–318, Siena, Italy, September 2004.

[3] B. Pesquet-Popescu and V. Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 3, pp. 1793–1796, Salt Lake City, Utah, USA, May 2001.

[4] A. Secker and D. Taubman, "Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression," *IEEE Transactions on Image Processing*, vol. 12, no. 12, pp. 1530–1542, 2003.

[5] M. Flierl and B. Girod, "Video coding with motion-compensated lifted wavelet transforms," *Signal Processing: Image Communication*, vol. 19, no. 7, pp. 561–575, 2004.

[6] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): design and construction," in *Proceedings of Data Compression Conference (DCC '99)*, pp. 158–167, Snowbird, Utah, USA, March 1999.

[7] S. S. Pradhan, J. Kusuma, and K. Ramchandran, "Distributed compression in a dense microsensor network," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 51–60, 2002.

[8] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): design and construction," *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 626–643, 2003.

[9] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1250–1276, 2002.

[10] J. Garcia-Frias, "Compression of correlated binary sources using turbo codes," *IEEE Communications Letters*, vol. 5, no. 10, pp. 417–419, 2001.

[11] A. D. Liveris, Z. Xiong, and C. N. Georghiades, "Compression of binary sources with side information at the decoder using LDPC codes," *IEEE Communications Letters*, vol. 6, no. 10, pp. 440–442, 2002.

[12] A. Aaron and B. Girod, "Compression with side information using turbo codes," in *Proceedings of Data Compression Conference (DCC '02)*, pp. 252–261, Snowbird, Utah, USA, April 2002.

[13] Y. Zhao and J. Garcia-Frias, "Data compression of correlated non-binary sources using punctured turbo codes," in *Proceedings of Data Compression Conference (DCC '02)*, pp. 242–251, Snowbird, Utah, USA, April 2002.

[14] Z. Xiong, A. D. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 80–94, 2004.

[15] M. Gastpar, P. L. Dragotti, and M. Vetterli, "The distributed Karhunen-Loève transform," in *Proceedings of IEEE Workshop on Multimedia Signal Processing (MMSP '02)*, pp. 57–60, St.Thomas, Virgin Islands, USA, December 2002.

[16] M. Gastpar, P. L. Dragotti, and M. Vetterli, "The distributed, partial, and conditional Karhunen-Loève transforms," in *Proceedings of Data Compression Conference (DCC '03)*, pp. 283–292, Snowbird, Utah, USA, March 2003.

[17] M. Gastpar, P. L. Dragotti, and M. Vetterli, "On compression using the distributed Karhunen-Loève transform," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 3, pp. 901–904, Montreal, Quebec, Canada, May 2004.

[18] M. Gastpar, P. L. Dragotti, and M. Vetterli, "The distributed Karhunen-Loève transform," submitted to *IEEE Transactions on Information Theory*.

[19] D. Rebollo-Monedero, A. Aaron, and B. Girod, "Transforms for high-rate distributed source coding," in *Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers (ACSSC '03)*, vol. 1, pp. 850–854, Pacific Grove, Calif, USA, November 2003.

[20] D. Rebollo-Monedero, S. D. Rane, and B. Girod, "Wyner-Ziv quantization and transform coding of noisy sources at high rates," in *Proceedings of the 38th Asilomar Conference on Signals, Systems and Computers (ACSSC '04)*, vol. 2, pp. 2084–2088, Pacific Grove, Calif, USA, November 2004.

[21] A. Aaron, S. D. Rane, E. Setton, and B. Girod, "Transform-domain Wyner-Ziv codec for video," in *Visual Communications and Image Processing (VCIP '04)*, vol. 5308 of *Proceedings of the SPIE*, pp. 520–528, San Jose, Calif, USA, January 2004.

[22] M. Flierl, "Distributed coding of dynamic scenes," Tech. Rep. EPFL-ITS-2004.015, Swiss Federal Institute of Technology, Lausanne, Switzerland, January 2004, http://lts1pc19.epfl.ch/repository/Flierl2004_780.pdf.

[23] M. Flierl and P. Vandergheynst, "Video coding with motion-compensated temporal transforms and side information," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 5, pp. 921–924, Philadelphia, Pa, USA, March 2005, invited paper.

[24] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing*, M. Landy and J. A. Movshon, Eds., pp. 3–20, MIT Press, Cambridge, Mass, USA, 1991.

[25] N. Gehrig and P. L. Dragotti, "Distributed compression in camera sensor networks," in *Proceedings of IEEE 6th Workshop on Multimedia Signal Processing (MMSP '04)*, pp. 311–314, Siena, Italy, September 2004.

[26] N. Gehrig and P. L. Dragotti, "Distributed compression of the plenoptic function," in *Proceedings of IEEE International Conference on Image Processing (ICIP '04)*, vol. 1, pp. 529–532, Singapor, October 2004.

[27] S. S. Pradhan and K. Ramchandran, "Enhancing analog image transmission systems using digital side information: a new wavelet-based image coding paradigm," in *Proceedings of the*

Data Compression Conference (DCC '01), pp. 63–72, Snowbird, Utah, USA, March 2001.

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

[29] X. Zhu, A. Aaron, and B. Girod, "Distributed compression for large camera arrays," in *Proceedings of IEEE Workshop on Statistical Signal Processing (SSP '03)*, pp. 30–33, St. Louis, Mo, USA, September 2003.

[30] R. Puri and K. Ramchandran, "PRISM: an uplink-friendly multimedia coding paradigm," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 4, pp. 856–859, Hong Kong, China, April 2003.

[31] A. Aaron, R. Zhang, and B. Girod, "Wyner-Ziv coding of motion video," in *Proceedings of the 36th Asilomar Conference on Signals, Systems and Computers (ACSSC '02)*, vol. 1, pp. 240–244, Pacific Grove, Calif, USA, November 2002.

[32] M. Flierl and B. Girod, "Video coding with motion compensation for groups of pictures," in *Proceedings of IEEE International Conference on Image Processing (ICIP '02)*, vol. 1, pp. 69–72, Rochester, NY, USA, September 2002.

[33] M. Flierl and B. Girod, *Video Coding with Superimposed Motion-Compensated Signals: Applications to H.264 and Beyond*, Kluwer Academic, Boston, Mass, USA, 2004.

[34] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1971.

Markus Flierl is with the Max Planck Center for Visual Computing and Communication at Stanford University, California. He is heading the Max Planck Research Group on Visual Sensor Networks. His research interests include visual data compression, information processing, sensor networks, multiview imaging, and motion in image sequences. He received his Dipl.-Ing. degree in electrical engineering as well as his Doctoral degree from Friedrich Alexander University, Erlangen, Germany, in 1997 and 2003, respectively. From 1999 to 2001, he was a scholar with the Graduate Research Center at Friedrich Alexander University. From 2000 to 2002, he joined the Information Systems Laboratory at Stanford University as a Visiting Researcher. From 2003 to 2005, he was a Postdoctoral Researcher with the Signal Processing Institute at the Swiss Federal Institute of Technology Lausanne, Switzerland. During his doctoral research, he contributed to the ITU-T Video Coding Experts Group standardization efforts on H.264. He is also a coauthor of an internationally published monograph.

Pierre Vandergheynst received the M.S. degree in physics and the Ph.D. degree in mathematical physics from the Université Catholique de Louvain, Louvain, Belgium, in 1995 and 1998, respectively. From 1998 to 2001, he was a Postdoctoral Researcher with the Signal Processing Laboratory, Swiss Federal Institute of Technology (EPFL) Lausanne, Switzerland. He is now an Assistant Professor of visual information representation theory at EPFL, where his research focuses on computer vision, image and video analysis, and mathematical techniques for applications in visual information processing. He is Co-Editor-in-Chief of EURASIP Journal on Signal Processing.