VIDEO CODING WITH MOTION-COMPENSATED TEMPORAL TRANSFORMS AND SIDE INFORMATION

Markus Flierl and Pierre Vandergheynst

Signal Processing Institute
Swiss Federal Institute of Technology, Lausanne
markus.flierl@epfl.ch, pierre.vandergheynst@epfl.ch

ABSTRACT

We address the problem of coding video signals in the presence of correlated video side information. The correlated video signals may originate from cameras that monitor the same scene from different view points. We utilize motion-compensated temporal wavelet transforms to decorrelate the video signals in temporal direction. One camera signal will provide the side information to improve the coding of the second camera signal. The side information is utilized in the transform domain after disparity compensation in the image domain. Given the correlation of the video side information, we investigate the bit rate reduction for the video coding scheme. Interestingly, the efficiency of multi-view side information is dependent on the level of temporal decorrelation: For a given correlation-SNR of the side information, bit rate savings due to side information are decreasing with improved temporal decorrelation.

1. INTRODUCTION

It is well known that video signals can be compressed more efficiently if correlated video side information is available at encoder and decoder. For example, two video cameras monitor the same scene from different view points. Therefore, the camera signals are correlated and the two encoders can exploit the dependency to improve the overall rate distortion performance. In one compression scenario, both encoders communicate with each other and compress the two video signals jointly. In an alternative compression scenario, both encoders do not communicate with each other but rely solely on the joint decoding of the video signals. A special case of the latter is source coding with side information. Wyner and Ziv 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 [1]. 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.

This paper discusses in particular stereo video coding with motion-compensated temporal wavelet transforms and side information. Both the video signal to be coded and the side information are image sequences captured form the same scene. The video signals are encoded with a motion-compensated lifted wavelet transform [2, 3, 4]. The side information is utilized in the transform domain after disparity compensation in the image domain.

Based on a signal model, we investigate the efficiency of motion-compensated temporal transform coding with side information. We extend the model for motion-compensated temporal transform coding [4, 5] and incorporate the impact of the side information. We assume very accurate disparity compensation to generate the side information. Further, we discuss performance bounds and compare to coding without side information.

The paper is organized as follows: Section 2 outlines our video coding scheme that utilizes video side information. We discuss the used motion-compensated wavelet transform as well as the coding of the quantized transform coefficients. We conclude with experimental results. Section 3 explores the efficiency of video coding with side information. With a model for transform-coded video signals, we determine the temporal conditional Karhunen-Loeve transform and discuss bounds for the coding gain due to side information.

2. CODING SCHEME UTILIZING VIDEO SIDE INFORMATION

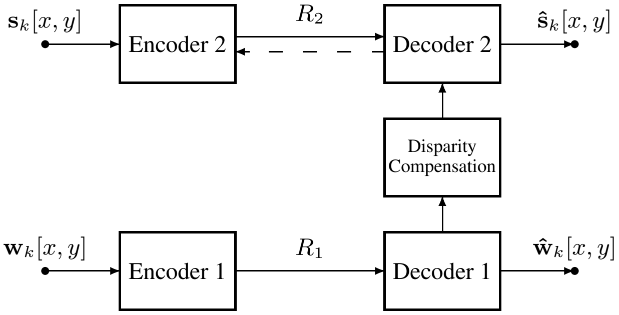

Fig. 1. Scheme for video coding with side information.

Fig. 1 depicts the scheme for video coding with side information. The two image sequences are represented by $\mathbf{s}_k[x,y]$ and $\mathbf{w}_k[x,y]$. The coding scheme comprises *Encoder 1* and *Encoder 2* that are not directly connected as well as *Decoder 2* that is dependent on *Decoder 1*. The side information for *Decoder 2* is improved by disparity compensation. As the video signals are not stationary, *Decoder 2* is decoding with feed-back.

2.1. Motion-Compensated Transform and Coefficient Coding

Each encoder in Fig. 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 [2]. The temporal transform provides K output pictures that are decomposed by a spatial 8×8 DCT. Details on the adaptive transform part of the coding scheme are given in [4].

Encoder 1 in Fig. 1 encodes the side information for *Decoder 2*. 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* uses nested lattice codes to represent the DCT coefficients of all temporal bands such that *Decoder 2* is able to exploit the side information efficiently.

Consider the 64 transform coefficients $\mathbf{c}_i$ of the 8×8 DCT at *Encoder 2*. The correlation between the i-th transform coefficient $\mathbf{c}_i$ at *Encoder 2* and the i-th 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 [6]. As we use uniform scalar quantization, we consider the 1-dimensional lattice. Let the fine code be $\mathcal{C}_0$ in the Euclidean space with minimum distance Q. $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$ are nested codes with the ν-th 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 $\mathcal{C}_0$, i.e. $\bigcup_\nu \mathcal{C}_{1,\nu} = \mathcal{C}_0$.

The binary representation of the quantized transform coefficients determines its coset representation in the nested lattice. If the transmission rate for a coefficient is $R_{TX} = \mu$, then the μ least significant bits of the binary representation determine the ν-th coset $\mathcal{C}_{\mu,\nu}$. For highly correlated coefficients, the number of required cosets and, hence, the transmission rate is small. To achieve efficient entropy coding of the binary representation of all 64 transform coefficients, we define bit planes. Each bit plane is run-length encoded and transmitted to *Decoder 2* upon request.

2.2. Decoding with Disparity Compensated Side Information

At *Encoder 2*, the quantized transform coefficients are represented with bit planes. *Encoder 2* is able to provide the full bit planes, independent of any side information at the *Decoder 2*. *Encoder 2* is also able to receive a bit plane mask to weight the current bit plane. The masked bit plane is run-length encoded and transmitted to *Decoder 2*.

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

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

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

With increasing number of received bit planes, i.e. increasing transmission rate $R_{TX}[i]$, this estimate gets more accurate and stays definitely constant for rates beyond the critical transmission rate $R^*_{TX}[i]$. 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 improve the efficiency of *Decoder 2*, the side information from *Decoder 1* is disparity 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 bit rate.

2.3. Experimental Results

For the experiments, we select the stereoscopic sequence *Tunnel* in QCIF resolution. We divide each view into groups of K 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.

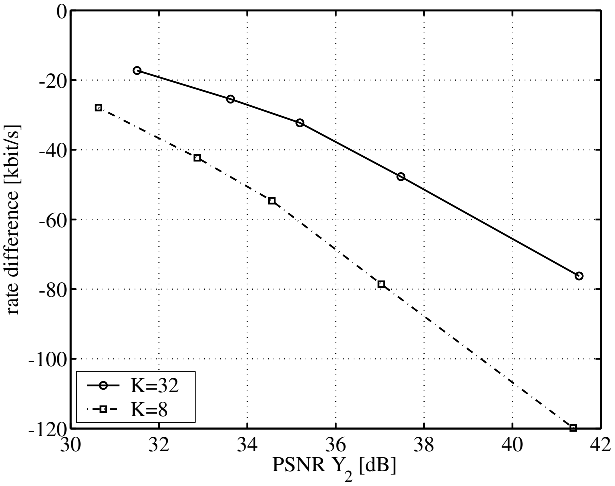

Fig. 2. Bit rate difference vs. luminance PSNR at *Decoder 2* for the sequence *Tunnel 2* (right view). The rate difference is the bit rate for decoding with side information minus the bit rate for decoding without side information and reflects the bit rate savings due to decoding with side information. Smaller bit rate savings are observed for strong temporal decorrelation (K=32) when compared to the bit rate savings for weak temporal decorrelation (K=8).

Fig. 2 shows the bit rate difference between decoding with side information and decoding without side information over the luminance PSNR at *Decoder 2* for the sequence *Tunnel 2* (right view). The bit rate 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. For example, at a quality of 35 dB, the bit rate is reduced from 360 to 330 kbit/s for $K = 32$, and from 470 to 410 kbit/s for $K = 8$. The total bit rate ranges between 200 and 1000 kbit/s for this experiment. It is observed that strong temporal filtering results in lower bit rate savings due to side information when compared to the bit rate savings due to side information for weaker temporal filtering. Obviously, there is a trade-off between the level of temporal decorrelation and the efficiency of multi-view side information.

3. EFFICIENCY OF VIDEO CODING WITH SIDE INFORMATION

In the following, we outline a signal model to study video coding with side information in more detail. We derive performance bounds and compare to coding without 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 [4, 5]. Let $\mathbf{s}_k = \{\mathbf{s}_k[x,y], (x,y) \in \Pi\}$ be scalar random fields over a two-dimensional orthogonal grid Π with horizontal and vertical spacing of 1. In Fig. 3, 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 $\mathbf{n}_k$ [4]. $\boldsymbol{\Delta}_k$ is the displacement error in the k-th picture, statistically independent from the model picture $\mathbf{v}$ and the noise $\mathbf{n}_k$ but correlated to other displacement errors. We assume a 2-D normal distribution with variance $\sigma^2_{\boldsymbol{\Delta}}$ and zero mean where the x- and y-components are statistically independent.

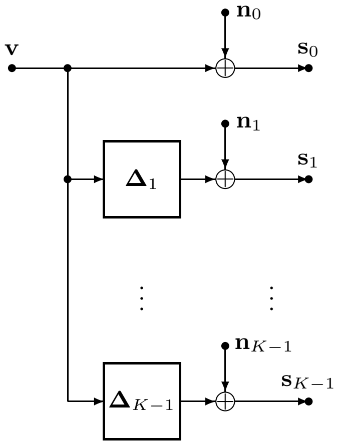

Fig. 3. Signal model for a group of K pictures.

From [4], 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.

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

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

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

$P = P(\omega)$ is the characteristic function of the continuous 2-D Gaussian displacement error.

$$P(\omega) = E\left\{e^{-j\omega^T\boldsymbol{\Delta}_k}\right\} = e^{-\frac{1}{2}\omega^T\omega\sigma^2_{\boldsymbol{\Delta}}} \qquad (4)$$

3.2. Rate Distortion with Video Side Information

Now, we consider the video coding scheme in Fig. 1 at high rates such that the reconstructed side information approaches the original side information $\hat{\mathbf{w}}_k \rightarrow \mathbf{w}_k$. With that, we have a Wyner-Ziv scheme (Fig. 4) and the rate distortion function R^* of *Encoder 2* is bounded by the conditional rate distortion function [1].

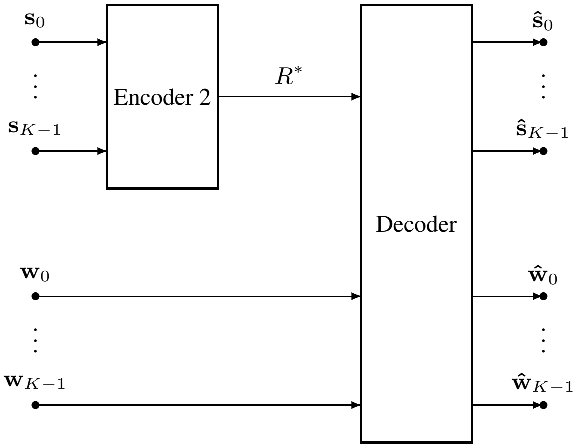

Fig. 4. Coding of K pictures $\mathbf{s}_k$ at rate R^* with side information of K pictures $\mathbf{w}_k$ at the decoder.

In the following, we assume very accurate disparity compensation and consider only illumination changes. We model the side information as a noisy version of the video signal to be encoded, i.e. $\mathbf{w}_k = \mathbf{s}_k + \mathbf{u}_k$, and assume that the noise $\mathbf{u}_k$ is also Gaussian with variance $\sigma^2_{\mathbf{u}}$ and independent of $\mathbf{s}_k$. 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 power spectral densities of the side information noise $\Phi_{\mathbf{uu}}(\omega) = \gamma(\omega)\Phi_{\mathbf{vv}}(\omega)I$. $\gamma = \gamma(\omega)$ 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 $\mathbf{v}$.

$$\gamma(\omega) = \frac{\Phi_{\mathbf{u}_k\mathbf{u}_k}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} \quad \text{for} \quad k = 0, 1, \ldots, K-1 \qquad (5)$$

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 $\mathbf{w}_k$.

$$\Phi_{\mathbf{s}|\mathbf{w}}(\omega) = \Phi_{\mathbf{ss}}(\omega) - \Phi^H_{\mathbf{ws}}(\omega)\Phi^{-1}_{\mathbf{ww}}(\omega)\Phi_{\mathbf{ws}}(\omega) \qquad (6)$$

With the model in Section 3.1 and the assumptions in Section 3.2, we obtain for the normalized conditional spectral density matrix

$$\frac{\Phi_{\mathbf{s}|\mathbf{w}}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} = \frac{1+\alpha-P}{1+\alpha+\gamma-P}\gamma I + \frac{P}{1+\alpha+\gamma-P}\cdot\frac{\gamma}{1+\alpha+\gamma+[K-1]P}\gamma\mathbf{1}\mathbf{1}^T. \quad (7)$$

For our signal model, the conditional KLT 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:

$$\begin{aligned}\frac{\Lambda_0^*(\omega)}{\Phi_{\mathbf{vv}}(\omega)} &= \frac{1+\alpha+\frac{\gamma KP}{1+\alpha+\gamma+[K-1]P}-P}{1+\alpha+\gamma-P}\gamma \\ \frac{\Lambda_k^*(\omega)}{\Phi_{\mathbf{vv}}(\omega)} &= \frac{1+\alpha-P}{1+\alpha+\gamma-P}\gamma \quad k=1,2,\dots,K-1 \quad (8)\end{aligned}$$

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 $\gamma(\omega) \to \infty$.

$$\begin{aligned}\frac{\Lambda_0^*(\omega)}{\Lambda_0(\omega)} &= \frac{\gamma}{1+\alpha+\gamma-P}\cdot\frac{1+\alpha+\frac{\gamma KP}{1+\alpha+\gamma+[K-1]P}-P}{1+\alpha+[K-1]P} \\ \frac{\Lambda_k^*(\omega)}{\Lambda_k(\omega)} &= \frac{\gamma}{1+\alpha+\gamma-P} \quad k=1,2,\dots,K-1 \quad (9)\end{aligned}$$

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

$$\Delta R_k^* = \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\frac{1}{2}\log_2\left(\frac{\Lambda_k^*(\omega)}{\Lambda_k(\omega)}\right)d\omega \quad (10)$$

It represents the maximum bit rate 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 [5].

Fig. 5 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 $\beta = \log_2(\sqrt{12}\sigma_{\mathbf{\Delta}}) = -1$. 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 bit rate savings are achievable if the GOP size K is smaller. The experimental results in Fig. 2 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.

4. CONCLUSIONS

This paper discusses coding of image sequences with video side information. The video signal to be coded and the video side information are correlated. We investigate a video coding scheme with a motion-compensated temporal transform that exploits the side information in the transform domain. At the encoder, the transform coefficients are coded with nested lattice codes, and at the decoder, the available side information is utilized for coset estimation. Regarding the coding efficiency, we observe a trade-off between the level of temporal decorrelation and the efficiency of multi-view side information.

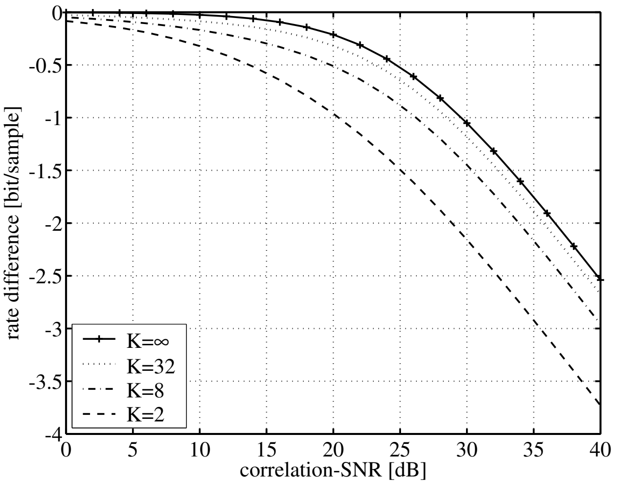

Fig. 5. Rate difference to motion-compensated transform coding without side information vs. correlation-SNR for groups of K pictures. The displacement inaccuracy β is -1 (half-pel accuracy) and the residual noise is -30 dB.

5. 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, pp. 1–10, Jan. 1976.

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

[3] 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, Dec. 2003.

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

[5] M. Flierl and B. Girod, "Video coding with motion compensation for groups of pictures," in *Proceedings of the IEEE International Conference on Image Processing*, Rochester, NY, Sept. 2002, vol. 1, pp. 69–72.

[6] 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, June 2002.