CODING EFFICIENCY OF VIDEO SENSOR NETWORKS

Markus Flierl

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

Invited Paper

ABSTRACT

We address the problem of coding multiple image sequences in a video sensor network. Multiple correlated video signals originate from cameras that monitor the same scene from different view points. Based on motion-compensated temporal transform coding of image sequences, the paper develops a high-rate approximation for the coding efficiency of a video sensor network with N cameras. We establish the rate difference between collaborative coding of correlated video signals and non-collaborative coding at each sensor. Bounds are obtained by assuming very accurate disparity compensation among all cameras. The goal is to study the impact of both inter-view correlation among video sensor signals and number of cameras in the network.

1. INTRODUCTION

This paper discusses N video sensors that monitor a non-static scene from different views. Each sensor comprises a video camera and a motion-compensated encoder. All video sensors are networked to a central decoder. As all video sensors sample the same scene from different views, the camera signals are correlated. This correlation can be exploited for the rate distortion efficiency of the video sensor network.

Video signals can be compressed more efficiently if correlated video side information is available at encoder and decoder. In one compression scenario, N encoders communicate with each other and compress the video signals jointly. In an alternative compression scenario, N 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 video sensor networks 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.

The encoder at each video sensor utilizes a motion-compensated temporal transform. In [2], motion-compensated temporal wavelets based on lifting implementations have been proposed. Motion-compensated temporal transforms permit open-loop video coding schemes that are rate distortion efficient. In [3], their rate distortion efficiency is investigated by modeling a motion-compensated subband coding scheme for a group of K pictures with a signal model for K motion-compensated pictures that are decorrelated by a linear transform. The Karhunen-Loeve transform is used to obtain performance bounds at high bit rates.

A transform-based approach to distributed source coding in image sensor networks seems promising. [4] discusses a framework for the distributed compression of vector sources. Each terminal applies a suitable local transform to its observation and encodes the resulting components separately in a Wyner-Ziv fashion, i.e., treating the compressed description of all other terminals as side information available to the decoder. [5] investigates Wyner-Ziv quantization and transform coding at high rates.

The outline of the paper is as follows: Section 2 draws the architecture of our video sensor network and discusses the coding problem. Section 3 investigates the rate distortion efficiency based on a model for transform-coded video signals.

2. CODING IN VIDEO SENSOR NETWORKS

Our video sensor network consists of N synchronized cameras, each providing K successive pictures. Each sensor encodes a motion-compensated version of its K camera images and transmits the data to a central decoder. The central decoder receives N data streams and reconstructs NK camera images. Each camera captures a different view point of the same scene. We assume that each camera image can be efficiently modeled by a disparity-compensated reference image. Signal components that cannot be modeled are captured by an additive model error. This model error shall be orthogonal to the disparity-compensated signal and statistically independent white noise is used. Further, we assume that the position of each camera is exactly known. Hence, we are able to perform very accurate disparity compensation. Note that inaccurate disparity compensation causes a degradation in coding performance. To obtain the theoretical efficiency bounds, we assume very accurate disparity compensation.

Fig. 1 depicts encoding and decoding in a video sensor network with N sensors. The ν-th image sequence is represented by K successive pictures $\mathbf{s}_{(\nu-1)K+k}[x, y]$, where $\nu = 1, 2, \dots, N$ and $k = 1, 2, \dots, K$. x and y denote the horizontal and vertical pixel position in the picture, respectively. Each sensor transmits a data stream at rate R_ν to the central decoder. The decoder reconstructs NK pictures $\hat{\mathbf{s}}_i[x, y]$, $i = 1, 2, \dots, NK$. The distortion in the video sensor network is the expected value of the mean square error between camera and reconstructed images. For the following investigation, the sensor network operates at high bit rates such that the reconstructed images at the decoder approach the corresponding camera images.

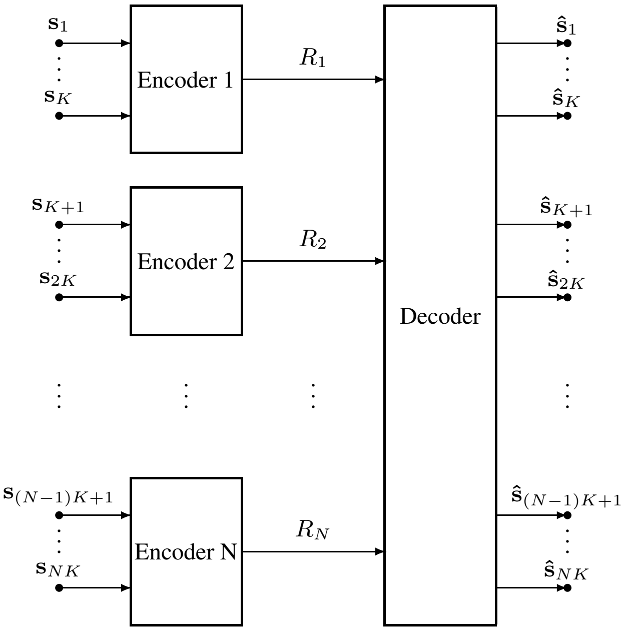

Fig. 1. Coding of N video sensor signals. Each image sequence ν, $\nu = 1, 2, \ldots, N$, consists of K pictures $\mathbf{s}_{(\nu-1)K+k}$, $k = 1, 2, \ldots, K$, and is transmitted at rate R_ν. The decoder reconstructs NK pictures $\hat{\mathbf{s}}_i$, $i = 1, 2, \ldots, NK$.

2.1. Coding of One Video Signal with Side Information

First, we consider only the ν-th video sensor and regard the remaining $N-1$ video signals as side information which enhance the coding efficiency of *Encoder* ν. The K images of the ν-th video signal are denoted by $\mathbf{u}_1, \ldots, \mathbf{u}_K$. $\mathbf{w}_1, \ldots, \mathbf{w}_{(N-1)K}$ are the remaining $N-1$ video signals.

As the video sensor network in Fig. 1 is studied at high rates, the reconstructed side information also approaches the original side information, i.e., $\hat{\mathbf{w}}_i \longrightarrow \mathbf{w}_i$ for $i = 1, \ldots, (N-1)K$. With that, we have a Wyner-Ziv scheme (Fig. 2) where the source $\mathbf{u}$ is encoded with *Encoder* ν and decoded in the presence of the side information $\mathbf{w}$. In this case, the rate distortion function R_ν^* of *Encoder* ν is bounded by the conditional rate distortion function [1].

As already pointed out, we assume very accurate disparity compensation. Additive white Gaussian noise $\mathbf{z}_{\mu,k}$, $\mu = 1, \ldots, N-1$ and $k = 1, \ldots, K$ is used for the model error. Therefore, the $N-1$ side information signals are noisy versions of the video signal $\mathbf{u}_k$, $k = 1, \ldots, K$, to be encoded, i.e., $\mathbf{w}_{(\mu-1)K+k} = \mathbf{u}_k + \mathbf{z}_{\mu,k}$, $\mu = 1, \ldots, N-1$. The noise $\mathbf{z}_{\mu,k}$ has variance $\sigma_{\mathbf{z}}^2$ for all μ, k, is mutually statistically independent as well as statistically independent of $\mathbf{u}_k$. With these assumptions, the rate distortion function R_ν^* of *Encoder* ν is equal to the conditional rate distortion function for coding $\mathbf{u}$ given the side information $\mathbf{w}$ [1].

The conditional rate distortion function enables us to investigate the efficiency of our video sensor network. Therefore, the conditional power spectral density matrix of the video signal $\mathbf{u}$ given the video side information $\mathbf{w}$ is of interest.

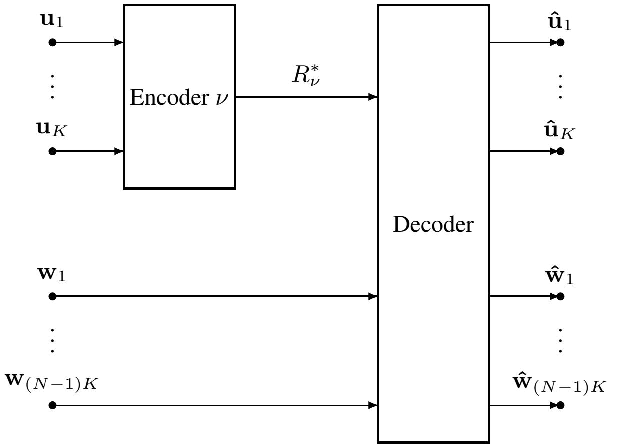

Fig. 2. Coding of K pictures $\mathbf{u}_1, \ldots, \mathbf{u}_K$ of sensor ν at rate R_ν^* with side information of $(N-1)K$ pictures $\mathbf{w}_1, \ldots, \mathbf{w}_{(N-1)K}$ at the decoder.

Let the cross spectral density of picture $\mathbf{u}_k[l]$ and $\mathbf{u}_\kappa[l]$, $k, \kappa = 1, 2, \ldots, K$, be denoted by $\Phi_{\mathbf{u}_k\mathbf{u}_\kappa}(\omega)$. Let $\Phi_{\mathbf{uu}}(\omega)$ be the power spectral density matrix of the K pictures $\mathbf{u}_k$ whose elements are $\Phi_{\mathbf{u}_k\mathbf{u}_\kappa}(\omega)$. For all remaining sensors $\mu = 1, \ldots, N-1$, let $\Phi_{\mathbf{zz}}(\omega)$ be the power spectral density matrix of K noise signals $\mathbf{z}_{\mu,k}$. With that, the power spectral density matrix of the side information is

$$\Phi_{\mathbf{ww}} = \begin{pmatrix} \Phi_{\mathbf{uu}} + \Phi_{\mathbf{zz}} & \Phi_{\mathbf{uu}} & \cdots & \Phi_{\mathbf{uu}} \\ \Phi_{\mathbf{uu}} & \Phi_{\mathbf{uu}} + \Phi_{\mathbf{zz}} & \cdots & \Phi_{\mathbf{uu}} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{\mathbf{uu}} & \Phi_{\mathbf{uu}} & \cdots & \Phi_{\mathbf{uu}} + \Phi_{\mathbf{zz}} \end{pmatrix}.$$

Please note the block structure of this matrix which permits us to write it as a sum of direct products

$$\Phi_{\mathbf{ww}} = \mathrm{I} \otimes \Phi_{\mathbf{zz}} + \mathbf{1}\mathbf{1}^T \otimes \Phi_{\mathbf{uu}}, \tag{1}$$

where I is the $(N-1) \times (N-1)$ identity matrix, and $\mathbf{1}\mathbf{1}^T$ is the $(N-1) \times (N-1)$ matrix with all elements equal to 1.

The cross spectral density matrix $\Phi_{\mathbf{wu}}(\omega)$ is also characterized by a block structure such that we can write the direct product

$$\Phi_{\mathbf{wu}}(\omega) = \mathbf{1} \otimes \Phi_{\mathbf{uu}}, \tag{2}$$

where $\mathbf{1}$ is the vector of length $N-1$ with all elements equal to 1.

With the well known conditional power spectral density matrix $\Phi_{\mathbf{u|w}}(\omega)$ of the video signal $\mathbf{u}$ given the video side information $\mathbf{w}$, i.e., $\Phi_{\mathbf{u|w}}(\omega) = \Phi_{\mathbf{uu}}(\omega) - \Phi_{\mathbf{wu}}^H(\omega)\Phi_{\mathbf{ww}}^{-1}(\omega)\Phi_{\mathbf{wu}}(\omega)$, we apply our assumptions in (1) and (2) and obtain the conditional power spectral density matrix that characterizes the video sensor network with N cameras:

$$\Phi_{\mathbf{u|w}}(\omega) = \Phi_{\mathbf{uu}}(\omega)\left[(N-1)\Phi_{\mathbf{uu}}(\omega) + \Phi_{\mathbf{zz}}(\omega)\right]^{-1}\Phi_{\mathbf{zz}}(\omega) \tag{3}$$

Interestingly, the power spectral density of the video signal $\mathbf{u}$ is weighted by $N-1$ when compared to the power spectral density of the inter-view correlation noise $\mathbf{z}$. That is, in a video sensor network with N cameras, the variance of the inter-view correlation noise is reduced by a factor of $N-1$. For a very large number of cameras N, the conditional power spectral density matrix is dominated by the reduced inter-view correlation noise.

$$\Phi_{\mathbf{u|w}}(\omega) \rightarrow \frac{\Phi_{\mathbf{zz}}(\omega)}{N-1} \quad \text{for} \quad N \rightarrow \infty \tag{4}$$

2.2. Coding of N Video Signals

So far, we have considered only the ν-th video sensor and have regarded the remaining $N-1$ video signals as side information which enhance the coding efficiency of *Encoder* ν. Now, we discuss the total bit rate which arrives at the central decoder.

We assume that the power spectral density of each camera signal is the same before disparity compensation. The signal of the current sensor always serves as a reference view point for disparity compensation. Therefore, only $N-1$ noisy versions of the current video signal are considered as side information. By changing the current sensor, we change the reference view point and, hence, adapt the disparity compensation. As the camera positions are known exactly, disparity compensation can be performed very accurately, independent of the current sensor. But we assume that the model error has the same variance, independent of the current sensor / reference view point. Consequently, each sensor shows the same rate distortion performance such that the total bit rate which arrives at the central decoder is N times the bit rate of the current sensor.

In the following, we will discuss the bit rate per video sensor for comparison purpose. But we keep in mind that the total bit rate that is received by the decoder is N times larger.

3. EFFICIENCY OF VIDEO SENSOR NETWORKS

In this section, we outline a video signal model to study the efficiency of collaborative coding in video sensor networks.

3.1. Model for Transform-Coded Video Signals

We build upon a model for motion-compensated subband coding of video that is outlined in [3]. A group of K pictures (GOP) is motion compensated with respect to a reference picture. The motion-compensated pictures are transform coded and the Karhunen-Loeve transform is used to obtain high-rate performance bounds. Any of the K pictures can be used as reference picture for motion compensation. We assume that we know the exact displacement information for all pictures relative to the reference frame but we permit also a displacement error to model inaccurate motion compensation. Further, statistically independent white Gaussian noise is added to model both occlusions due to motion and illumination changes.

[3] assumes that the pictures $\mathbf{u}_k$ are shifted versions of the model picture $\mathbf{v}$ which are also degraded by independent additive white Gaussian noise $\mathbf{n}_k$. The displacement error $\boldsymbol{\Delta}_k$ in the k-th picture is 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.

From [3], we adopt the power spectral density matrix of the pictures $\mathbf{u}_k$ 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 elements equal to 1.

$$\frac{\Phi_{\mathbf{uu}}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} = [1+\alpha(\omega)-P(\omega)]\,\mathrm{I} + P(\omega)\mathbf{1}\mathbf{1}^T \tag{5}$$

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

$$\alpha(\omega) = \frac{\Phi_{\mathbf{n}_k\mathbf{n}_k}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} \quad \text{for} \quad k=1,2,\ldots,K \tag{6}$$

$P = P(\omega) = \exp(-\frac{1}{2}\omega^T\omega\sigma^2_{\boldsymbol{\Delta}})$ is the characteristic function of the continuous 2-D Gaussian displacement error.

3.2. Conditional Power Spectral Density Matrix

Given the power spectral density matrix $\Phi_{\mathbf{uu}}(\omega)$ in (5), we determine the conditional power spectral density matrix in (3) for our model assumptions. In Section 2.1, we assumed that the inter-view correlation noise $\mathbf{z}_{\mu,k}$ is mutually statistically independent. Therefore, its power spectral density matrix is given by

$$\Phi_{\mathbf{zz}}(\omega) = \gamma(\omega)\Phi_{\mathbf{vv}}(\omega)\mathrm{I}, \tag{7}$$

where I is the $K\times K$ identity matrix and $\gamma = \gamma(\omega)$ is the normalized power spectral density of the side information noise $\Phi_{\mathbf{z}_k\mathbf{z}_k}(\omega)$ with respect to the model picture $\mathbf{v}$.

$$\gamma(\omega) = \frac{\Phi_{\mathbf{z}_k\mathbf{z}_k}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} \quad \text{for} \quad k=1,2,\ldots,K \tag{8}$$

Finally, we obtain the normalized $K\times K$ conditional power spectral density matrix for our model assumptions as follows

$$\frac{\Phi_{\mathbf{u}|\mathbf{w}}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} = \frac{Q}{[N-1]Q+\gamma}\gamma\mathrm{I} + \frac{P}{[N-1]Q+\gamma}\cdot\frac{\gamma}{[N-1][Q+KP]+\gamma}\gamma\mathbf{1}\mathbf{1}^T, \tag{9}$$

where $Q = Q(\omega) = 1+\alpha(\omega)-P(\omega)$. Note that both I and $\mathbf{1}\mathbf{1}^T$ are $K\times K$ matrices.

3.3. Conditional Karhunen-Loeve Transform

Now, it is sufficient to use the conditional Karhunen-Loeve transform and code the N video signals at high rates in order to achieve the conditional rate distortion function.

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. Therefore, a set of N suitable bases can be chosen for our sensor network.

Finally, K eigendensities are needed to determine the performance bounds. They are obtained from (9) as follows:

$$\frac{\Lambda_1^*(\omega)}{\Phi_{\mathbf{vv}}(\omega)} = \frac{Q+\frac{\gamma KP}{[N-1][Q+KP]+\gamma}}{[N-1]Q+\gamma}\gamma, \tag{10}$$

$$\frac{\Lambda_k^*(\omega)}{\Phi_{\mathbf{vv}}(\omega)} = \frac{Q}{[N-1]Q+\gamma}\gamma, \quad k=2,3,\ldots,K. \tag{11}$$

3.4. Coding Gain for Video Sensor Networks

With the conditional eigendensities (10) and (11), we are able to determine the coding gain due to collaborative coding in video sensor networks. We normalize the conditional eigendensities $\Lambda_k^*(\omega)$ with respect to the eigendensities $\Lambda_k(\omega)$ that we obtain for non-collaborative coding as $\Lambda_k^*(\omega)\to\Lambda_k(\omega)$ for $\gamma(\omega)\to\infty$. The normalized conditional eigendensities are

$$\frac{\Lambda_1^*(\omega)}{\Lambda_1(\omega)} = \frac{\gamma}{[N-1]Q+\gamma}\cdot\frac{Q+\frac{\gamma KP}{[N-1][Q+KP]+\gamma}}{Q+KP}, \tag{12}$$

$$\frac{\Lambda_k^*(\omega)}{\Lambda_k(\omega)} = \frac{\gamma}{[N-1]Q+\gamma}, \quad k=2,3,\ldots,K. \tag{13}$$

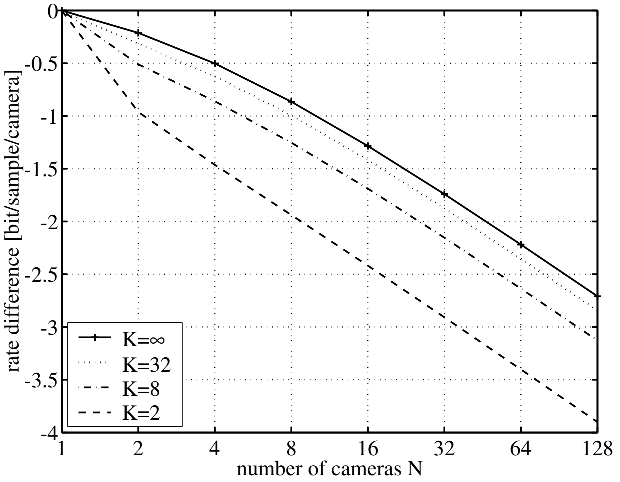

Fig. 3. Rate difference to non-collaborative motion-compensated transform coding vs. number of cameras N for groups of K pictures. The displacement inaccuracy β of motion compensation among K pictures is -1 (half-pel accuracy), the RNL is -30 dB, and the correlation-SNR is 20 dB.

The rate difference to non-collaborative coding is used to measure the improved compression efficiency for each picture k of the ν-th sensor.

$$\Delta R^*_{\nu,k} = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{1}{2} \log_2 \left(\frac{\Lambda^*_k(\omega)}{\Lambda_k(\omega)} \right) d\omega \qquad (14)$$

It represents the maximum bit rate reduction (in bit/sample) possible by optimum encoding of the eigensignal in the case of collaborative coding, compared to optimum encoding of the eigensignal without collaborative coding for Gaussian wide-sense stationary signals for the same mean square reconstruction error. The rate difference ΔR^*_ν of the ν-th sensor is the average over all K eigensignals. Note that the rate differences of all sensors are identical due to our model assumptions.

We plot the average rate difference to non-collaborative motion-compensated transform coding as a function of the number of cameras N, of the correlation-SNR c-SNR $= 10\log_{10}([1 + \sigma^2_{\mathbf{n}}]/\sigma^2_{\mathbf{z}})$, and of the displacement inaccuracy $\beta = \log_2(\sqrt{12}\sigma_{\mathbf{\Delta}})$. We choose half-pel accurate motion compensation, i.e., $\beta = -1$. For all graphs, the residual noise level RNL $= 10\log_{10}(\sigma^2_{\mathbf{n}})$ is -30 dB which is common for practical video sequences. Note that the variance of the model picture $\mathbf{v}$ is normalized to $\sigma^2_{\mathbf{v}} = 1$.

Fig. 3 depicts the average rate difference to non-collaborative coding over the number of cameras N. It investigates the GOP size K for motion-compensated transform coding at a correlation-SNR of 20 dB. For the comparison, collaborative and non-collaborative coding utilize the same GOP size K. Note that for a large number of cameras, the rate difference decreases by 0.5 bit per sample per camera if the number of cameras doubles.

Fig. 4 shows the average rate difference to non-collaborative coding over the correlation-SNR. It shows the impact of the number of cameras N in the sensor network for a GOP size of $K = 32$ pictures. For highly correlated video signals, the gain due to collaborative coding increases by 1 bit per sample if the c-SNR increases by 6 dB.

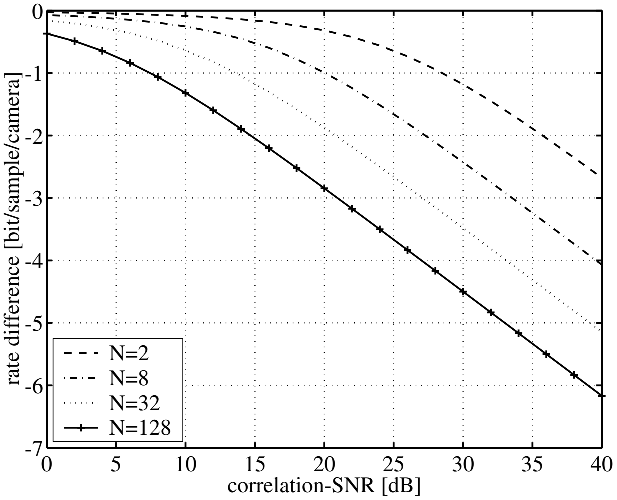

Fig. 4. Rate difference to non-collaborative motion-compensated transform coding vs. correlation-SNR for N cameras. The displacement inaccuracy β of motion compensation among $K = 32$ pictures is -1 (half-pel accuracy) and the RNL is -30 dB.

Finally, we point out that these results are bounded by the above assumptions. First, assuming independent white noise for model errors like occlusions and illumination differences is limiting. These errors may also be correlated. Second, the high-rate assumption guarantees a high quality of the side information. Quantization noise in the side information will degrade its efficiency. Third, the Wyner-Ziv bound may not be achieved if the innovation is not Gaussian. Fourth, numerical results are obtained with Gaussian signals. Actual video can be compressed more efficiently.

4. CONCLUSIONS

We discussed coding of multiple image sequences in a video sensor network. We observe with our model that, first, for a large number of cameras, the rate difference to non-collaborative motion-compensated coding decreases by 0.5 bit per sample per camera if the number of cameras doubles. Second, for highly correlated sensor signals, the gain due to collaborative coding increases by 1 bit per sample if the correlation-SNR increases by 6 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] 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.

[4] M. Gastpar, P. Dragotti, and M. Vetterli, "On compression using the distributed Karhunen-Loève transforms," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Montreal, Canada, May 2004, vol. III, pp. 901–904.

[5] D. Rebollo-Monedero, S. Rane, and B. Girod, "Wyner-Ziv quantization and transform coding of noisy sources at high rates," in *Proceedings of the Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, Nov. 2004.