Motion and Disparity Compensated Coding for Video Camera Arrays

Markus Flierl, Aditya Mavlankar, and Bernd Girod *

Information Systems Laboratory
Department of Electrical Engineering, Stanford University
Stanford, CA 94305

Abstract. Video camera arrays are used to capture multi-view imagery of dynamic 3D scenes. To communicate a scene to a remote location, multiple video signals have to be compressed. Disparity compensation exploits the correlation among the image sequences and may improve the rate-distortion efficiency of the communication system. Motion compensation makes use of the temporal correlation within each image sequence and may also improve the rate-distortion efficiency. For both cases, the accuracy of compensation determines the efficiency. We study the impact on the overall rate-distortion efficiency of both disparity compensation and motion compensation. Further, we discuss the benefit of coding across the views given the number of temporal frames used for decorrelation.

Index Terms: multi-view image sequences, motion and disparity compensated coding, video camera arrays.

1 Introduction

Capturing dynamic scenes can be accomplished with a video camera array. Such an array may be part of a three-dimensional TV system which enables users to view a distant 3D world freely [1]. A critical component of such systems is the coding engine that compresses the multi-view video data into a rate-distortion efficient representation. The most straightforward approach to the multi-view coding problem is to temporally encode the individual video streams independent of one another [2]. But efficient coding can be achieved by exploiting the correlation in temporal direction as well as the correlation among the views.

Temporal correlation may be exploited by motion compensation between temporally successive pictures of each video camera. Disparity compensation between neighboring camera views may take advantage of correlation across the views. To study the impact of the accuracy of compensation, the high-rate model for video coding with motion-compensated lifted wavelet transforms in [3] is extended to assess the efficiency of coding multi-view video sequences. Further, the impact of jointly encoding N views as well as K temporally successive pictures of each view-sequence is also investigated. These model results are compared to data obtained from coding experiments with selected multi-view sequences.

The paper is organized as follows: Section 2 outlines the investigated experimental coding scheme using hierarchical and generalized B pictures and presents the obtained coding results. Section 3 discusses a mathematical model for multi-view video coding and establishes performance bounds based on optimal transform coding.

2 Coding Scheme

The coding scheme encodes jointly a Matrix of Pictures (MOP) with N image sequences, where each consists of a group of K temporally successive pictures. Each MOP is encoded with one I frame and $NK - 1$ hierarchical and generalized B frames.

2.1 Hierarchical and Generalized B Frames

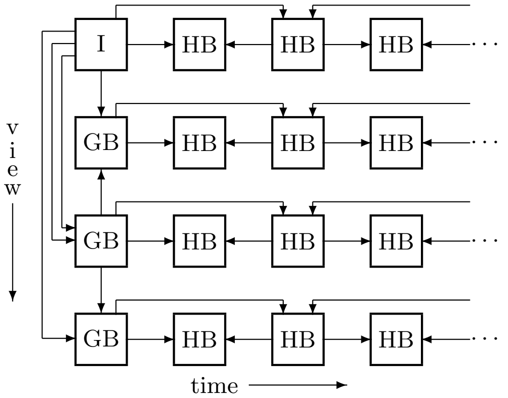

Fig. 1. Matrix of pictures (MOP) for $N = 4$ image sequences, each comprising of a group of $K = 4$ temporally successive pictures and its encoding.

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

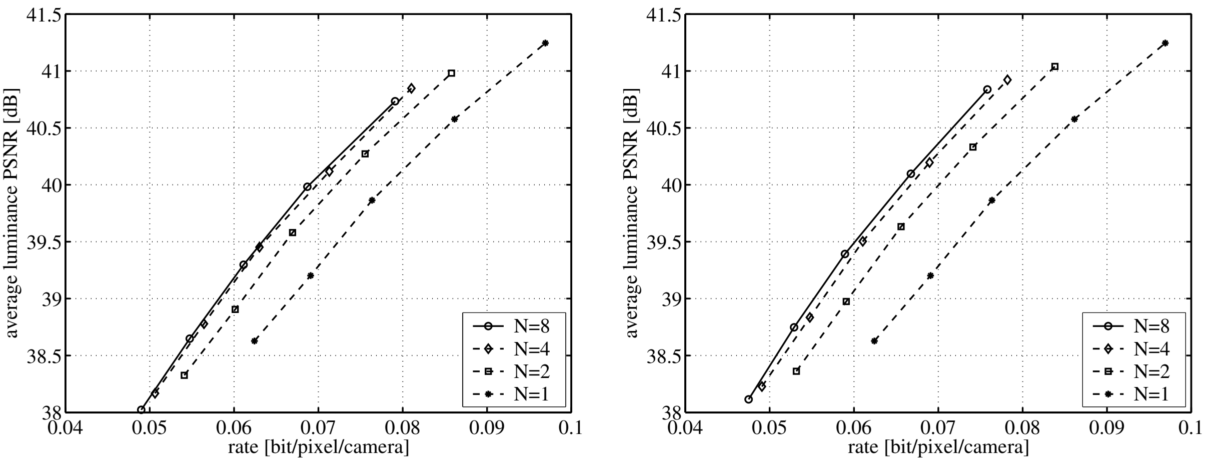

Fig. 2. Average luminance PSNR vs. bit-rate for encoding 8 view-sequences of *Ballet*. The performance is plotted for a GOV size of $N = 1, 2, 4$, and 8. The disparity compensation is integer-pel accurate **(left)** and quarter-pel accurate **(right)**. The temporal GOP size is $K = 8$ and motion compensation is quarter-pel accurate.

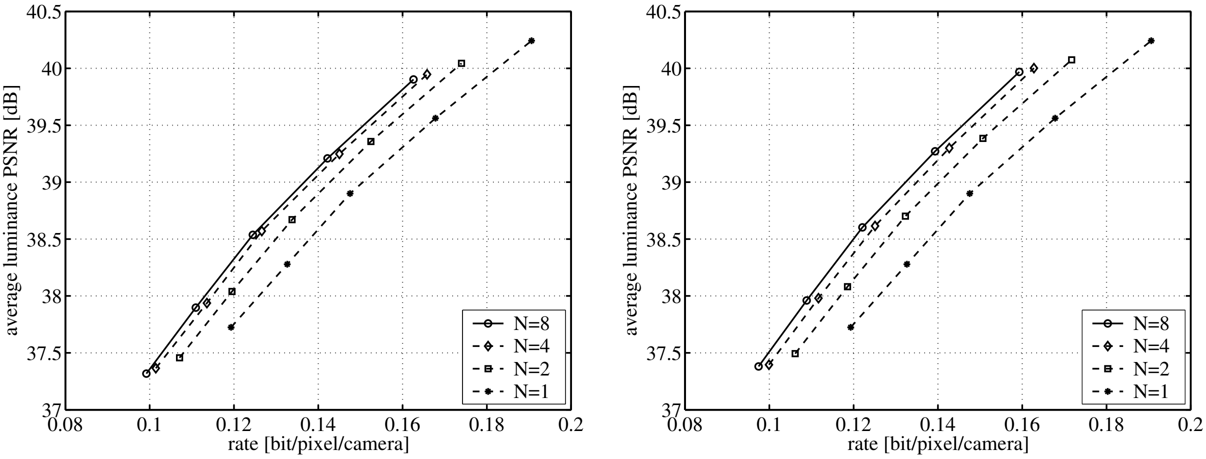

Fig. 3. Average luminance PSNR vs. bit-rate for encoding 8 view-sequences of *Breakdancers*. The performance is plotted for a GOV size of $N = 1, 2, 4$, and 8. The disparity compensation is integer-pel accurate **(left)** and quarter-pel accurate **(right)**. The temporal GOP size is $K = 8$ and motion compensation is quarter-pel accurate.

We use the hierarchical and generalized B frames of H.264 to implement the encoding of a MOP. At every K-th time instant, we encode N view images with one I frame and $N - 1$ hierarchical B frames. As bi-directional prediction is not always possible, we use generalized B frames (GB) with bi-predictive coding as depicted in **Fig. 1**. The reconstructed N view images at every K-th time instant are now used as reference for the hierarchical B frames (HB) in temporal direction. This encoding permits view scalability as temporal B frames of the current view have no reference to neighboring view sequences.

We consider the hierarchical B frames as an approximation of a dyadic wavelet decomposition in view as well as in time direction. As we process the view direction first, followed by the time direction only, we consider this encoding as an approximation for a separable decomposition.

2.2 Experimental Results

We use the multi-view sequences *Ballet* and *Breakdancers* each with 8 views and a resolution of 256 × 192. For the encoding, we choose the same quantizer parameter for all pictures of a MOP.

The first experiment investigates the impact of accurate disparity compensation. **Fig. 2** and **Fig. 3** depict the performance of *Ballet* and *Breakdancers*,

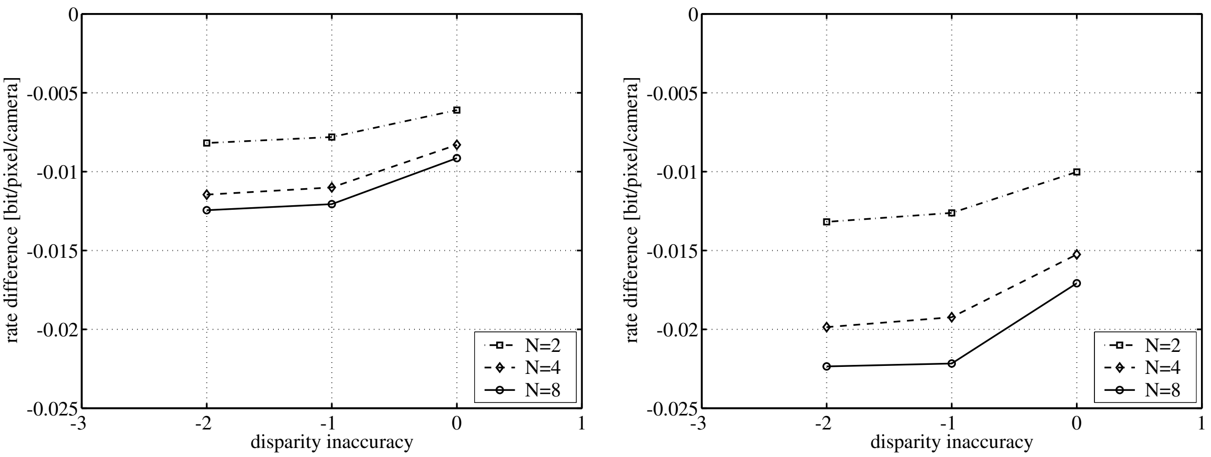

Fig. 4. Rate difference to independent encoding of each camera signal vs. disparity inaccuracy of disparity compensation for 8 view-sequences of *Ballet* **(left)** and of *Breakdancers* **(right)**. The performance is plotted for a GOV size of $N = 2, 4$, and 8, where $N = 1$ is the reference. The temporal GOP size is $K = 8$ and motion compensation is quarter-pel accurate. The rates are obtained for PSNR = 40 dB.

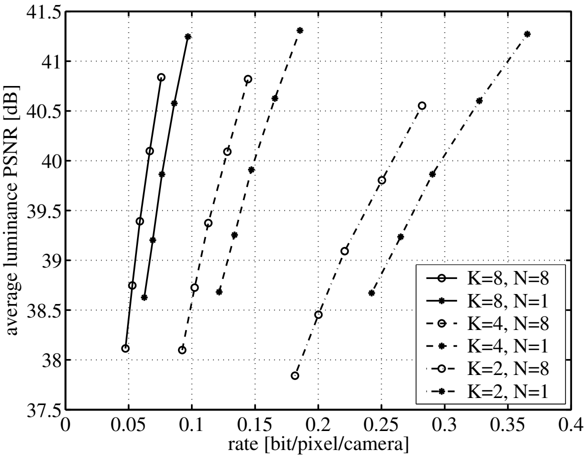

Fig. 5. Average luminance PSNR vs. bit-rate for encoding 8 view-sequences of *Ballet*. The performance is plotted for a GOV size of $N = 1$ and 8 as well as a temporal GOP size of $K = 2, 4$, and 8. Both disparity and motion compensation are quarter-pel accurate.

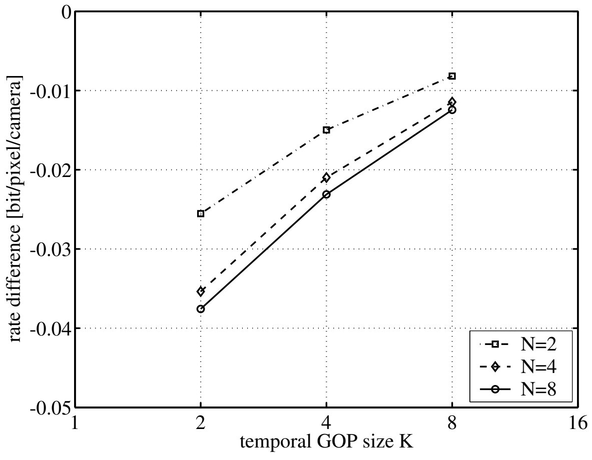

Fig. 6. Rate difference to independent encoding of each camera signal vs. temporal GOP size K for 8 view-sequences of *Ballet*. The performance is plotted for a GOV size of $N = 2, 4$, and 8, where $N = 1$ is the reference. Both disparity and motion compensation are quarter-pel accurate. The PSNR is 40 dB.

respectively, for various sizes of the Group of Views (GOV) N. The left plot in each figure shows the performance of integer-pel, the right plot that of quarter-pel accurate disparity compensation.

To study the rate difference to independent encoding of each camera signal, we choose the case $N = 1$ as the reference and plot the rate difference in **Fig. 4** at a PSNR of 40 dB. Note that the rate difference is the actual rate minus the rate for independent encoding of each camera signal. Hence, it is negative if the coding efficiency improves over the reference. We observe that the efficiency improves when increasing the accuracy from integer-pel (0) to half-pel (-1) and quarter-pel (-2). The improvement due to accurate compensation is larger if we perform disparity compensation among $N = 8$ views when compared to compensation among $N = 2$ only.

The second experiment investigates the impact of the temporal GOP size K on the overall coding performance for *Ballet*. **Fig. 5** shows the rate distortion performance and **Fig. 6** plots the rate difference at a PSNR of 40 dB. The coding scheme with a temporal GOP size of $K = 8$ and GOV size of $N = 8$ shows a much smaller improvement over its reference scheme with $K = 8$ and $N = 1$ than the coding scheme with a temporal GOP size of $K = 2$ and GOV size of $N = 8$ over its reference scheme with $K = 2$ and $N = 1$. This effect gets weaker for smaller GOV size N.

3 Mathematical Model for Multi-View Video Coding

To study the previous observations more closely, we outline a signal model that shall capture the effects of accurate motion and disparity compensation as well as the dimensions of the MOP on the coding efficiency. We extend the signal model for K motion-compensated pictures in [3] to a model for NK disparity- and motion-compensated pictures. These pictures are then decorrelated by the Karhunen-Loeve Transform (KLT) for optimal encoding and for achieving rate distortion bounds.

3.1 Signal Model

The model assumes that multiple view-sequences are generated from a model image sequence which is shifted by a disparity error vector $\mathbf{\Theta} = (\mathbf{\Theta}_x, \mathbf{\Theta}_y)^T$ and distorted by additive white Gaussian noise $\mathbf{z}$. The shift shall model disparity compensation with limited accuracy and the noise shall capture signal components that cannot be modeled by a translatory disparity. Further, it is assumed that the model image sequence $\{\mathbf{c}_k, k = 1, 2, \dots, K\}$ with power spectral density matrix $\Phi_{\mathbf{cc}}(\omega)$ is generated from a model picture $\mathbf{v}$ with power spectral density (PSD) $\Phi_{\mathbf{vv}}(\omega)$, which is shifted by a displacement error vector $\mathbf{\Delta}_{1k} = (\mathbf{\Delta}_{x,1k}, \mathbf{\Delta}_{y,1k})^T$ and distorted by additive white Gaussian noise $\mathbf{n}_k$. **Fig. 7** summarizes the model. Note that all K temporal pictures of the ν-th view, $\nu = 1, 2, \dots, N$, are shifted by the disparity vector $\mathbf{\Theta}_{1\nu}$, where the reference view is the first view.

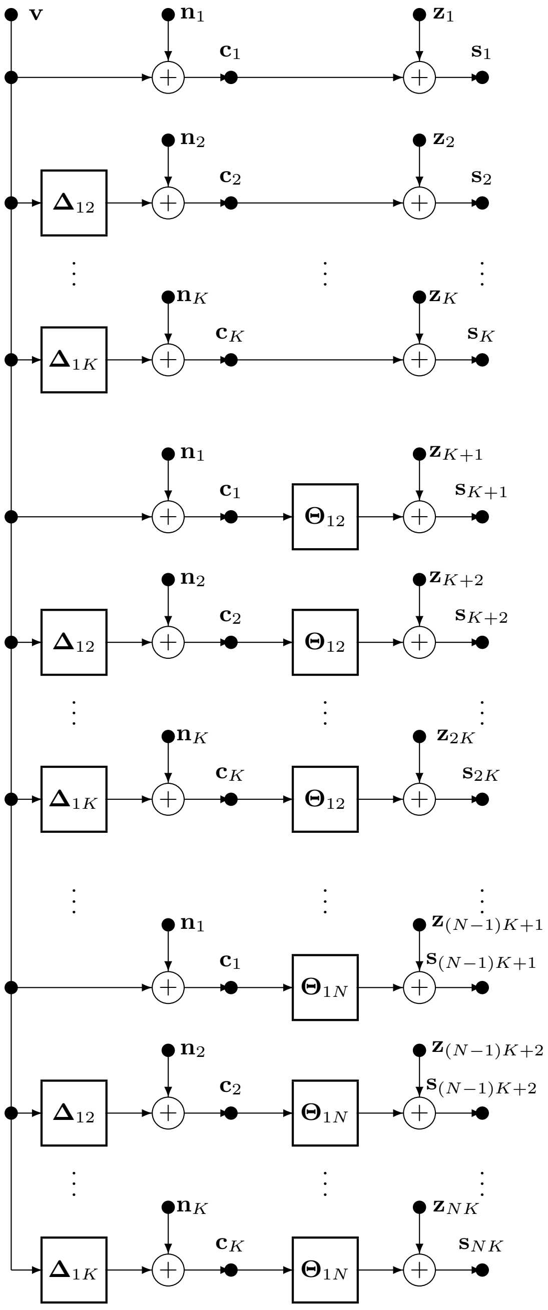

Fig. 7. Signal model for N image sequences each comprising of a group of K temporally successive pictures.

[3] assumes the principle of additive motion for the true motion in the sequence, i.e., $\mathbf{d}_{\kappa\mu} + \mathbf{d}_{\mu\nu} = \mathbf{d}_{\kappa\nu}$, as well as for the estimated motion, i.e., $\hat{\mathbf{d}}_{\kappa\mu} + \hat{\mathbf{d}}_{\mu\nu} = \hat{\mathbf{d}}_{\kappa\nu}$. Consequently, the principle of additive motion holds also for the displacement error $\mathbf{\Delta}_{\kappa\mu} + \mathbf{\Delta}_{\mu\nu} = \mathbf{\Delta}_{\kappa\nu}$. In the following, we assume also additive disparity, and consequently, additive disparity error $\mathbf{\Theta}_{\kappa\mu} + \mathbf{\Theta}_{\mu\nu} = \mathbf{\Theta}_{\kappa\nu}$. Further, we assume that any temporal picture can be the temporal reference picture. This implies that the variances of all displacement errors are identical. Similarly, any view can be a reference view and the variances of all disparity errors are identical. Finally, we consider displacements and disparities as mutually statisti-

cally independent. Hence, we use mutually statistically independent displacement and disparity errors for the model.

We adopt from [3] the PSD matrix of the model image sequence, normalized to the PSD of the model picture.

$$\frac{\Phi_{\mathbf{cc}}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} = \begin{pmatrix} 1+\alpha(\omega) & P(\omega) & \cdots & P(\omega) \\ P(\omega) & 1+\alpha(\omega) & \cdots & P(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ P(\omega) & P(\omega) & \cdots & 1+\alpha(\omega) \end{pmatrix} \quad (1)$$

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

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

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

$$P(\omega) = E\left\{e^{-j\omega^T \mathbf{\Delta}_{\mu\nu}}\right\} = e^{-\frac{1}{2}\omega^T\omega\sigma_{\mathbf{\Delta}}^2} \quad (3)$$

With the signal model in **Fig. 7** and the above assumptions for the displacement and disparity errors, the PSD matrix of N view-sequences is

$$\frac{\Phi_{\mathbf{ss}}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} = \Gamma(\omega) \otimes \frac{\Phi_{\mathbf{cc}}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} + \mathrm{I}\gamma(\omega), \quad (4)$$

where $\otimes$ denotes the Kronecker product, I the $NK \times NK$ identity matrix, and $\Gamma(\omega)$ the characteristic matrix of the disparity error.

$$\Gamma(\omega) = \begin{pmatrix} 1 & G(\omega) & \cdots & G(\omega) \\ G(\omega) & 1 & \cdots & G(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ G(\omega) & G(\omega) & \cdots & 1 \end{pmatrix} \quad (5)$$

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

$$G(\omega) = E\left\{e^{-j\omega^T \mathbf{\Theta}_{\mu\nu}}\right\} = e^{-\frac{1}{2}\omega^T\omega\sigma_{\mathbf{\Theta}}^2} \quad (6)$$

Finally, $\gamma(\omega)$ is the normalized power spectral density of the view noise $\Phi_{\mathbf{z}_i\mathbf{z}_i}(\omega)$ with respect to the model picture $\mathbf{v}$.

$$\gamma(\omega) = \frac{\Phi_{\mathbf{z}_i\mathbf{z}_i}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} \quad \text{for} \quad i = 1, 2, \ldots, NK \quad (7)$$

Note that the PSD matrix of N view-sequences can be written as a Kronecker product between the characteristic matrix $\Gamma(\omega)$ and the PSD matrix of the model image sequence as we assume mutual statistical independence between displacement and disparity errors.

3.2 Transform Coding Gain

Now, we determine the performance bound by optimal transform coding with the KLT. For that, we determine the eigenvalues of the PSD matrix $\Phi_{\mathbf{ss}}(\omega)$ in (4). Note that the eigenvalues of a matrix resulting from a Kronecker product are simply the Kronecker product of the eigenvalues of the individual factors. The eigenvalues of $\Phi_{\mathbf{cc}}(\omega)$ are $\lambda_1(\omega) = 1+\alpha(\omega)+(K-1)P(\omega)$ and $\lambda_2(\omega) = 1+\alpha(\omega)-P(\omega)$. The eigenvalues of $\Gamma(\omega)$ are $\lambda_3(\omega) = 1+(N-1)G(\omega)$ and $\lambda_4(\omega) = 1-G(\omega)$. Hence, the eigenvalues $\Lambda_i^*(\omega)$ of $\Phi_{\mathbf{ss}}(\omega)$ are:

$$\frac{\Lambda_i^*(\omega)}{\Phi_{\mathbf{vv}}(\omega)} = \begin{cases} \lambda_1(\omega)\lambda_3(\omega) + \gamma(\omega) : & 1\times \\ \lambda_1(\omega)\lambda_4(\omega) + \gamma(\omega) : & (N-1)\times \\ \lambda_2(\omega)\lambda_3(\omega) + \gamma(\omega) : & (K-1)\times \\ \lambda_2(\omega)\lambda_4(\omega) + \gamma(\omega) : & (N-1)(K-1)\times \end{cases} \quad (8)$$

The reference coding scheme encodes the sequences independently and does not exploit the correlation across the N views. Hence, it encodes eigenvalues $\Lambda_i(\omega)$ as follows:

$$\frac{\Lambda_i(\omega)}{\Phi_{\mathbf{vv}}(\omega)} = \begin{cases} \lambda_1(\omega) + \gamma(\omega) : & N\times \\ \lambda_2(\omega) + \gamma(\omega) : & N(K-1)\times \end{cases} \quad (9)$$

Note that for both schemes the eigenvalues sum to $NK[1+\alpha(\omega)+\gamma(\omega)]\Phi_{\mathbf{vv}}(\omega)$.

We assess the performance of the multi-view coding scheme by using the average rate difference to independent encoding of N view-sequences.

$$\Delta R = \frac{1}{NK} \sum_{i=1}^{NK} \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{1}{2} \log_2 \frac{\Lambda_i^*(\omega)}{\Lambda_i(\omega)} d\omega \quad (10)$$

It represents the maximum bit rate reduction (in bit/sample/camera) possible by optimum encoding of the eigensignals in the case of joint coding, compared to optimum encoding of the eigensignals for independent coding, for Gaussian wide-sense stationary signals for the same mean square reconstruction error [3].

In the following, we plot the average rate difference for GOV size N to independent coding of N view-sequences as a function of the temporal GOP size K as well as of the disparity inaccuracy $\vartheta = \log_2(\sqrt{12}\sigma_{\mathbf{\Theta}})$. For both graphs, the residual motion noise level motion-RNL $= 10\log_{10}(\sigma_{\mathbf{n}}^2)$ is -30 dB, which is common for practical video sequences. The residual view noise level view-RNL $= 10\log_{10}(\sigma_{\mathbf{z}}^2)$ is -10 dB reflecting a large disparity model error to capture new scene content. Note that $\sigma_{\mathbf{v}}^2 = 1$. The motion inaccuracy $\beta = \log_2(\sqrt{12}\sigma_{\mathbf{\Delta}})$ is a function of

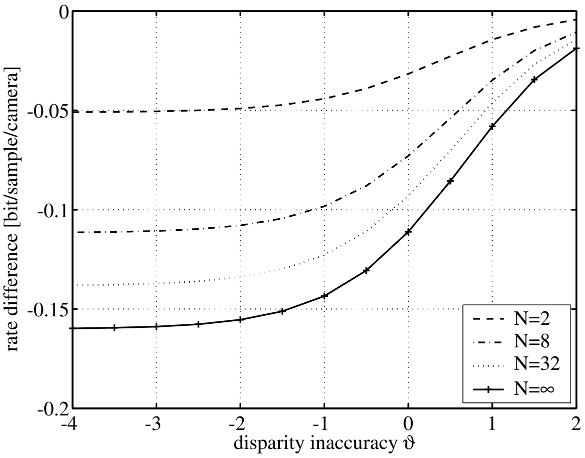

Fig. 8. Rate difference to independent encoding of each camera signal vs. disparity inaccuracy ϑ of disparity compensation for GOV sizes of N. The displacement inaccuracy β of motion compensation among $K = 8$ pictures is -2 (quarter-pel accuracy) and the motion-RNL is -30 dB. The view-RNL is -10 dB.

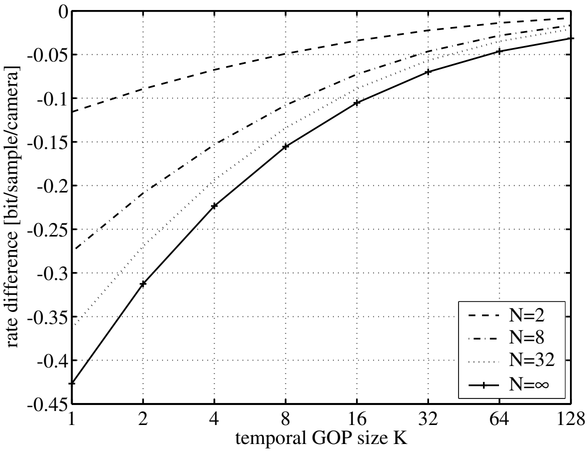

Fig. 9. Rate difference to independent encoding of each camera signal vs. temporal GOP size K for groups of N views. The displacement inaccuracy β of motion compensation among K pictures as well as the disparity inaccuracy ϑ of disparity compensation among N views is -2 (quarter-pel accuracy). The motion-RNL is -30 dB, the view-RNL is -10 dB.

the variance of the displacement error components $\sigma^2_{\mathbf{\Delta}}$. The value $\beta = 0$ represents integer-pel accuracy, $\beta = -1$ half-pel accuracy, $\beta = -2$ quarter-pel accuracy, etc. For the graphs, β is chosen to be -2.

Fig. 8 depicts the average rate difference to independent encoding of each camera signal over the disparity inaccuracy ϑ of disparity compensation for a temporal GOP size of $K = 8$. The disparity inaccuracy $\vartheta = \log_2(\sqrt{12}\sigma_{\mathbf{\Theta}})$ is a function of the variance of the disparity error components $\sigma^2_{\mathbf{\Theta}}$ to improve the readability of the graph. The value $\vartheta = 0$ represents integer-pel accuracy, $\vartheta = -1$ half-pel accuracy, $\vartheta = -2$ quarter-pel accuracy, etc. We observe that for each GOV size N the rate efficiency over independent encoding improves for more accurate disparity compensation. This improvement is larger if we perform disparity compensation among $N = 8$ views when compared to compensation among $N = 2$ only. The experimental results in **Fig. 4** match these observations.

Fig. 9 depicts the rate difference in bit per sample per camera to independent encoding of each camera signal vs. temporal GOP size K for various GOV sizes N. The displacement inaccuracy β of motion compensation among K pictures as well as the disparity inaccuracy ϑ of disparity compensation among N views is -2 (quarter-pel accuracy). We observe that the coding scheme with a temporal GOP size of $K = 8$ and GOV size of $N = 8$ shows a much smaller improvement over its reference scheme with $K = 8$ and $N = 1$ than the coding scheme with a temporal GOP size of $K = 2$ and GOV size of $N = 8$ over its reference scheme with $K = 2$ and $N = 1$. This effect gets weaker for smaller GOV size N. The experimental results in **Fig. 6** match these observations.

4 Conclusions

We study the problem of coding jointly N multiview video sequences experimentally and theoretically. For groups of N views, we discuss the impact of both disparity inaccuracy and temporal GOP size K on the overall rate distortion performance. In particular, we observe that increasing the temporal GOP size K reduces the coding gain over independent sequence encoding.

References

1. Tanimoto, M.: Free viewpoint television - FTV. In: Proceedings of the Picture Coding Symposium, San Francisco, CA (2004)
2. Vetro, A., Matusik, W., Pfister, H., Xin, J.: Coding approaches for end-to-end 3D TV systems. In: Proceedings of the Picture Coding Symposium, San Francisco, CA (2004)
3. Flierl, M., Girod, B.: Video coding with motion-compensated lifted wavelet transforms. Signal Processing: Image Communication **19** (2004) 561–575