Motion and Disparity Compensated Coding for Multiview Video

Markus Flierl, *Member, IEEE*, Aditya Mavlankar, *Student Member, IEEE*, and Bernd Girod, *Fellow, IEEE*

(Invited Paper)

***Abstract*—We investigate the rate-distortion efficiency of motion and disparity compensated coding for multiview video. Disparity compensation exploits the correlation among the view sequences and motion compensation makes use of the temporal correlation within each view sequence. We define a matrix of pictures with N view sequences, each with K temporally successive pictures. For experimental coding purposes, a scheme based on H.264/AVC is devised. We assess the overall rate-distortion efficiency for matrices of pictures of various dimensions (N, K). Moreover, we discuss the impact of inaccurate disparity compensation within a matrix of pictures. Finally, we propose and discuss a theoretical model for multiview video coding that explains our experimental observations. Performance bounds are presented for high rates.**

***Index Terms*—Motion and disparity compensated coding, multiview image sequences, video camera arrays.**

I. Introduction

TODAY'S advances in display and camera technology enable new applications for 3-D scene communication. In free viewpoint video, image sequences recorded simultaneously with multiple cameras are transformed into a special data representation that enables interactive 3-D navigation through the dynamic scene [1]. Driven by multiview video, autostereoscopic 3-D displays can produce both stereo and movement parallax with a small number of views [2]. Other researchers have experimented with densely packed camera arrays [3]. Large video camera arrays may also be part of a 3-D TV system which enables users to view a distant 3-D world freely [4]. Such systems may be based on a ray-space representation as used for free viewpoint television (FTV) [5], [6]. Others may utilize the concept of a multivideo-plus-depth data representation [7]. An overview of 3-D video and free viewpoint video is given in [8].

For coding and transmission of multiview video, statistical dependencies within the multiview imagery have to be exploited. The captured images are characterized by disparities between views and motion between temporally successive frames. These are important parameters of the dynamic scene. The problem of structure and motion estimation in multiview teleconferencing-type sequences is discussed in more detail in [9]. To achieve a good tradeoff between scene quality and bit-rate, disparity and motion among all the pictures has to be exploited efficiently. Usually, this is accomplished with either predictive or subband coding schemes that perform both disparity and motion compensation.

Predictive coding uses previously decoded pictures as references for predicting the current picture. Disparity-compensated view prediction exploits correlation among the views and uses concepts known from motion-compensated prediction [10]. Further, view synthesis prediction may offer additional benefits. With the help of depth maps, virtual views are synthesized from previously encoded views, and subsequently, used for predictive coding of the current view [11], [12]. It is also possible to consider special systems where, e.g., the user viewpoints are constrained to a line. Compression and rendering of such simplified dynamic light fields is discussed in [13] and [14].

As an alternative to motion and disparity compensated predictive coding, one can perform a motion and disparity adaptive subband decomposition of the multiview video signal, followed by quantization and entropy coding of the subband coefficients. For static light fields, disparity-compensated wavelets have been investigated for compression purposes [15], [16]. Also, schemes for multiview wavelet video coding have been devised [17], [18]. The inherent scalability of such wavelet decompositions is appealing.

The rate-distortion efficiency of multiview video coding is of great interest. For single-view video coding, theoretical performance bounds have been established for motion-compensated prediction [19] as well as motion-compensated subband coding [20]–[22]. Obviously, the simplest approach to multiview coding is to encode the individual video sequences independently [23]. But for efficient multiview video coding, the similarity among the views should also to be taken into account. In [24], we have proposed a mathematical model that captures both inter-view correlation and temporal correlation. This model is based on the high-rate model for motion-compensated subband coding of video [20]–[22]. In the following, we will discuss and study this model in more detail. In particular, we are interested in the impact of the accuracy of disparity compensation on the coding efficiency. Further, we explore the encoding of N views, each with K temporally successive pictures and its impact on the overall coding performance. Finally, these model results are compared to data obtained from actual coding experiments with selected multiview video sequences. To emphasize the experimental observations, this paper first presents the experimental results. Then, we study the observations in more detail with the help of the mathematical model.

Manuscript received January 20, 2007. This work was supported by the Max Planck Center for Visual Computing and Communication. This paper was presented in part at the Picture Coding Symposium, Beijing, China, April 2006. This paper was recommended by Guest Editor Y. He.

The authors are with the Max Planck Center, Stanford University, Stanford, CA 94305 USA (e-mail: mflierl@ieee.org).

Digital Object Identifier 10.1109/TCSVT.2007.903780

Rate-distortion efficient compression of static light fields has previously been investigated in [25]–[29]. This work did not consider video sequences but focused on interactive streaming of static light fields using predictive light field coding. The theoretical part of the work includes interpolation between views, arbitrary prediction structures, and in particular, a model to link the inaccuracy of the underlying scene geometry model to the inaccuracy of the disparity between images.

Currently, multiview video coding is also investigated by the *Joint Video Team* (JVT). MPEG is one partner of the team and has previously explored video-based rendering technology [30]. Now, the JVT is developing a *Joint Multiview Video Model* (JMVM) [31] which is based on the video coding standard ITU-T Recommendation H.264–ISO/IEC 14496–10 AVC [32]. The current JMVM proposes illumination change-adaptive motion compensation and prediction structures with hierarchical B pictures [10]. The JMVM uses the block-based coding techniques of H.264/AVC to exploit both temporal and view correlation within temporally successive pictures and neighboring views. Based on H.264/AVC, we have devised an experimental coding scheme to investigate the efficiency of motion and disparity compensated coding of multiview video.

This paper is organized as follows. Section II outlines the investigated experimental coding scheme using bipredictive slices of H.264/AVC and presents the obtained coding results. Section III discusses a statistical signal model for multiview video coding and establishes performance bounds based on optimal transform coding.

II. Motion and Disparity Compensated Coding

In the following, we devise a coding scheme for multiview video with motion and disparity compensation. To achieve the best rate-distortion performance, the statistical dependencies among all the pictures should be exploited. Towards this end, we arrange the multiview video data into a *matrix of pictures* (MOP). Each MOP consists of N image sequences, each with K temporally successive pictures. With that, we consider the correlation among all the pictures within a MOP. The MOP is then encoded jointly by our multiview video coding scheme.

Our scheme is based on the state-of-the-art video coding standard H.264/AVC. With this standard, generalized B pictures [33] are available. The concept is implemented in the form of *bipredictive slices*. Bipredictive slices may utilize a linear combination of any two motion-compensated signals for prediction. This improves coding efficiency, particularly for multihypothesis motion [34]. Moreover, they may themselves be used as a reference for further prediction. We use these two features of bipredictive slices to construct an efficient coding scheme for multiview video with H.264/AVC.

A. Coding With Bipredictive Slices

We use bipredictive slices of H.264/AVC to perform a multiresolution decomposition of the MOP. Multiresolution signal decompositions [35] offer several benefits. In particular, they generate signal decompositions that may permit efficient compression [36]. For our application, we desire a multiresolution decomposition in both time and view direction.

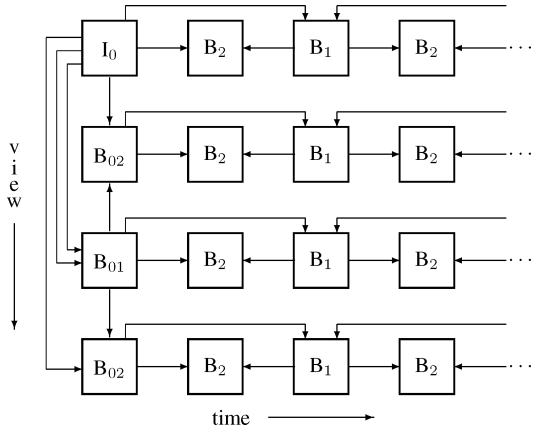

Fig. 1. Matrix of pictures (MOP) for $N = 4$ image sequences, each comprising of $K = 4$ temporally successive pictures. The coding structure with bipredictive slices is also shown.

We explain our multiresolution decomposition of the multiview video signal with the example in Fig. 1. It depicts a MOP for $N = 4$ image sequences, each comprising of $K = 4$ temporally successive pictures. Each MOP is encoded with one intra picture and $NK - 1$ generalized B pictures. First, each MOP is decomposed in view direction at the first time instant only. That is, the sequences have view decompositions at every Kth time instant. The intra picture I_0 in each MOP represents the lowest view resolution. The next view resolution level is attained by including the bipredictive slice B_{01}. The highest view resolution is achieved with the bipredictive slices B_{02}. Second, the reconstructed N view images at every Kth time instant are now used as reference for multiresolution decompositions with bipredictive slices in temporal direction. The decomposition in view direction at every Kth time instant represents already the lowest temporal resolution level. The next temporal resolution level is attained by including the bipredictive slices B_1. The highest temporal resolution is achieved with the bipredictive slices B_2. Thus, the hierarchical decomposition of the MOP with bipredictive slices generates a representation with multiple resolutions in time and view direction.

In general, we encode at every Kth time instant N view images with one I slice and $N - 1$ bipredictive slices. If possible, bipredictive slices may use bidirectional prediction, i.e., references from neighboring left and right views may be combined in case of horizontal camera arrangements. But rate-distortion optimal reference picture selection at the encoder determines the optimal reference pair. The reconstructed N view images at every Kth time instant are now used as reference for bipredictive slices in temporal direction. We perform hierarchical decompositions with bipredictive slices in time and view direction. This permits view scalability since temporal slices of each view have no reference to their neighboring view slices. Our scheme can be easily extended to incorporate disparity-compensated prediction at all time instances, as discussed in [10]. Such extensions come with a significant increase in complexity while offering only a limited overall coding gain.

Note that our hierarchical decomposition with bipredictive slices is similar to a dyadic wavelet decomposition in time direction, followed by a decomposition in view direction. However,

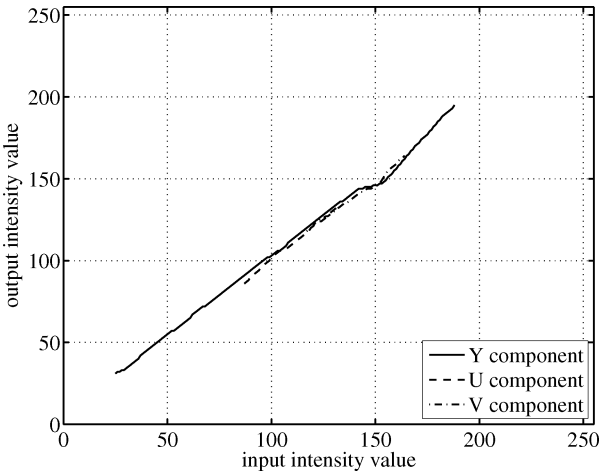

Fig. 2. Intensity value mapping for camera view 8 of the multiview data set *Ballet*. Histogram matching is applied to luminance as well as chrominance components with camera view 4 as reference.

unlike a subband coder, our coder operates in "closed-loop" manner, i.e., bipredictive slices are reconstructed (with quantization errors) first, before they serve as reference for further prediction.

B. Compensation of Inter-View Intensity Variations

Multiview video sequences recorded with multiple video cameras often show substantial luminance and chrominance variations among the views. Even when using cameras of the same make and model for all views, inaccurate camera calibration may be the cause. To compensate for these variations, we use histogram matching [37], [38] as a preprocessing step.

Histogram matching may improve disparity-compensated prediction, and hence, the compression efficiency of multiview video coding. A particular view is assumed to be the reference. The method first computes cumulative histograms of the reference and current view sequences. Then, the compensation is achieved by mapping each value of the current view sequence to a corrected value. The goal of the mapping is to make the cumulative histograms of the reference and current view sequences similar. With the same reference view, this is repeated for all remaining views. Note that the same mapping function is used for all time instances to maintain temporal coherence for efficient motion compensation.

An example of a typical mapping function is given in Fig. 2. Histogram matching is applied to luminance and chrominance components of the multiview data set *Ballet*. The mappings of the component intensity values of camera view 8 with camera view 4 as the reference are shown. Histogram matching is able to compensate for luminance and chrominance variations.

C. Experimental Results

With our hierarchical coding scheme and the previous preprocessing method, we investigate experimentally the efficiency of motion and disparity compensated coding of multiview video. We use the five multiview video data sets *Ballet*, *Ballroom*, *Breakdancers*, *Exit*, and *Race1*, each with eight views. The spatial resolution is 256×192 for *Ballet* and *Breakdancers* and 320×240 for the remaining data sets. We have reduced the original spatial resolution of the data sets with the MPEG downsampling filter [39] to lower the computational burden of the extensive simulations.

H.264/AVC allows us to choose freely the size of the reference picture buffer. To achieve high compression performance, we use a large buffer. For example, if the temporal GOP size is $K = 8$, the reference picture buffer can hold up to seven pictures. The buffer gets populated as more pictures in the temporal GOP are coded. Hence, pictures of the finest temporal resolution will benefit the most from the large reference buffer. The same holds for encoding in view direction.

In the following, we report on two experiments which assess compression efficiency of multiview video signals. The first investigates the impact of disparity compensation accuracy. The second explores the impact of the temporal GOP size K. Both experiments consider always all eight views of the data set, but choose various sizes (N, K) for the MOP. Each experiment comprises two steps. In the first step, we measure rate-distortion curves, particularly at high image quality. In the second step, we choose a PSNR of 40 dB and calculate the rate difference to independent encoding of each view sequence, i.e., to the case $N = 1$.

Fig. 3 depicts the rate-distortion points for *Breakdancers* for various sizes of the *group of views* (GOV) N. The left plot in Fig. 3 shows the results for integer-pel, the right plot that for quarter-pel accurate disparity compensation. The temporal GOP size is $K = 8$ and motion compensation is quarter-pel accurate. For *Breakdancers*, preprocessing with the histogram matching method did not provide any benefits. Therefore, we use the original data set for the simulations. To study the rate difference to independent encoding of each view sequence, we choose the case $N = 1$ as 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 view sequence. 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 $N = 2$ views only.

The experiment is repeated for *Ballet*. For this data set, preprocessing with histogram matching is advantageous. Fig. 5 depicts the rate-distortion points with (left) integer-pel and (right) quarter-pel accurate disparity compensation. Fig. 6 shows the rate difference for (left) the preprocessed data set and that of the (right) original data set. Again, efficiency improves with the accuracy of disparity compensation and increasing GOV size N. Note that preprocessing offers a small but consistent relative benefit as all three curves shift towards higher coding efficiency. Further results demonstrate also a consistent absolute benefit as the advantage is also observed for the reference $N = 1$.

We continue with the second experiment and explore the impact of the temporal GOP size K on the overall coding performance. Fig. 7 depicts the rate-distortion points for the eight view sequences of (left) histogram matched *Ballet* and (right) original *Exit*. The results are recorded for various MOP

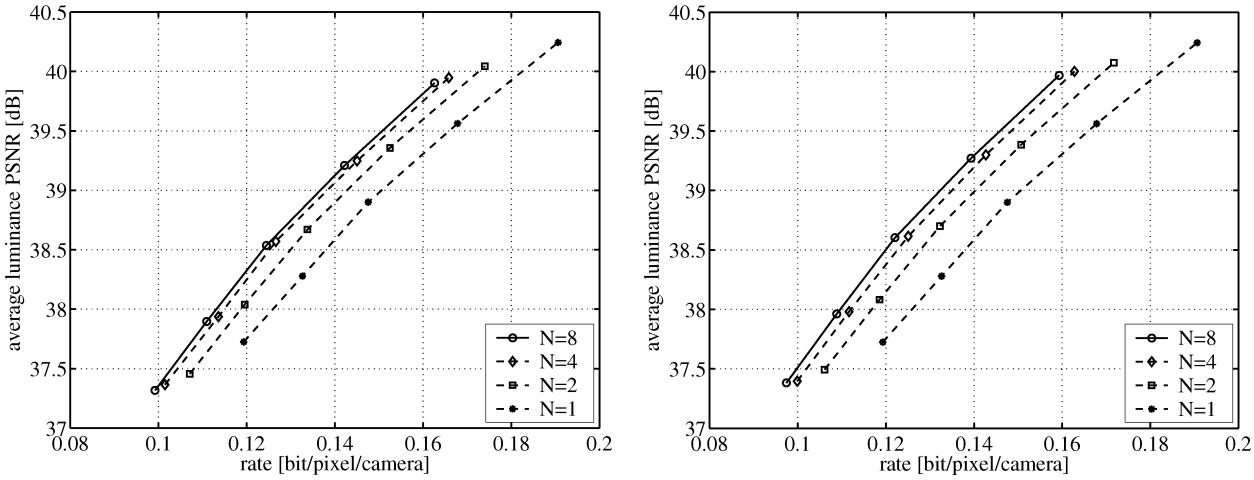

Fig. 3. Average luminance PSNR versus bit-rate for encoding 8 view-sequences of *Breakdancers*. The performance is plotted for a GOV size of $N = 1, 2, 4,$ and 8. The disparity compensation is (left) integer-pel accurate and (right) quarter-pel accurate. The temporal GOP size is $K = 8$ and motion compensation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

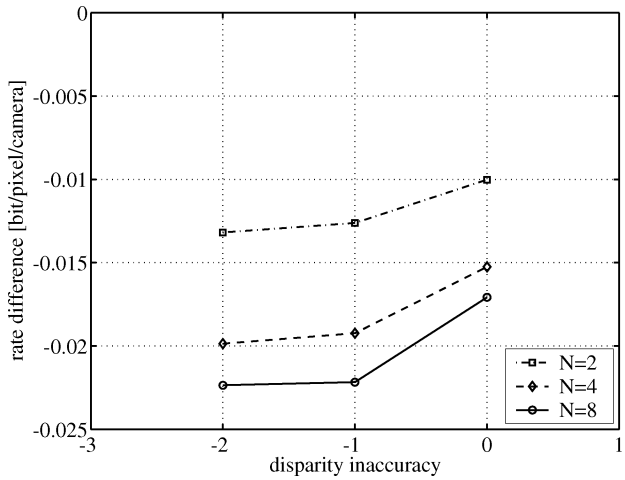

Fig. 4. Rate difference to independent encoding of each view sequence versus disparity inaccuracy of disparity compensation for 8 view-sequences of *Breakdancers*. The performance is plotted for a GOV size of $N = 2$, 4, and 8, where $N = 1$ is the reference. The temporal GOP size is $K = 8$ and motion compensation is quarter-pel accurate. The rates are obtained for $\mathrm{PSNR} = 40$ dB.

sizes (N, K). Both disparity and motion compensation are quarter-pel accurate. For *Exit*, preprocessing with the histogram matching method did not offer any benefits. Therefore, we use the original data set for the simulations. To study the average rate difference to independent encoding of each view sequence for various temporal GOP sizes K, we choose the case $N = 1$ as reference and plot the average rate differences at a PSNR of 40 dB. Fig. 8 shows the average rate difference for *Ballet* with (left) preprocessed and (right) original data set. Again, preprocessing offers a small but consistent relative benefit as all curves shift towards higher coding efficiency. Further, we observe that encoding with temporal GOP size of $K = 8$ and GOV size of $N = 8$ provides a much smaller improvement over its reference encoding with $K = 8$ and $N = 1$ than encoding with temporal GOP size of $K = 2$ and GOV size of $N = 8$ over its reference encoding with $K = 2$ and $N = 1$. This effect becomes weaker for smaller GOV size N.

This experiment is repeated for histogram matched *Ballroom* and original *Race1* (see Fig. 9) as well as for the original data sets *Breakdancers* and *Exit* (see Fig. 10). Again, the relative efficiency grows with increasing GOV size N. But with increasing temporal GOP size K, the relative efficiency grows slower with GOV size N.

Finally, we have observed that preprocessing with histogram matching does not always improve overall coding efficiency. When using the mapping function, histogram matching may introduce high spatial frequencies that deteriorate the overall coding performance. Also, histogram matching may affect the temporal coherence of the view sequences such that motion compensation is less efficient. But for the data sets *Ballet* and *Ballroom*, overall coding efficiency is improved.

III. Mathematical Model for Multiview Video Coding

To explain the previous observations, we outline a statistical signal model to capture the effects of motion compensation accuracy and disparity compensation accuracy as well as the dimensions of the MOP on the coding efficiency. We extend the signal model for K motion-compensated pictures in [22] 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.

A. Statistical Signal Model

The model assumes that multiple view sequences are generated from a *root image sequence* which is shifted by a disparity error vector $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_x, \boldsymbol{\Theta}_y)^T$ 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 root image sequence $\{\mathbf{c}_k, k = 1, 2, \ldots, K\}$ with power spectral density matrix $\Phi_{\mathbf{cc}}(\omega)$ is generated from a *root picture* $\mathbf{v}$ with power spectral density (PSD) $\Phi_{\mathbf{vv}}(\omega)$, which is shifted by a displacement error vector $\boldsymbol{\Delta}_{1k} = (\boldsymbol{\Delta}_{x,1k}, \boldsymbol{\Delta}_{y,1k})^T$ and distorted by additive white

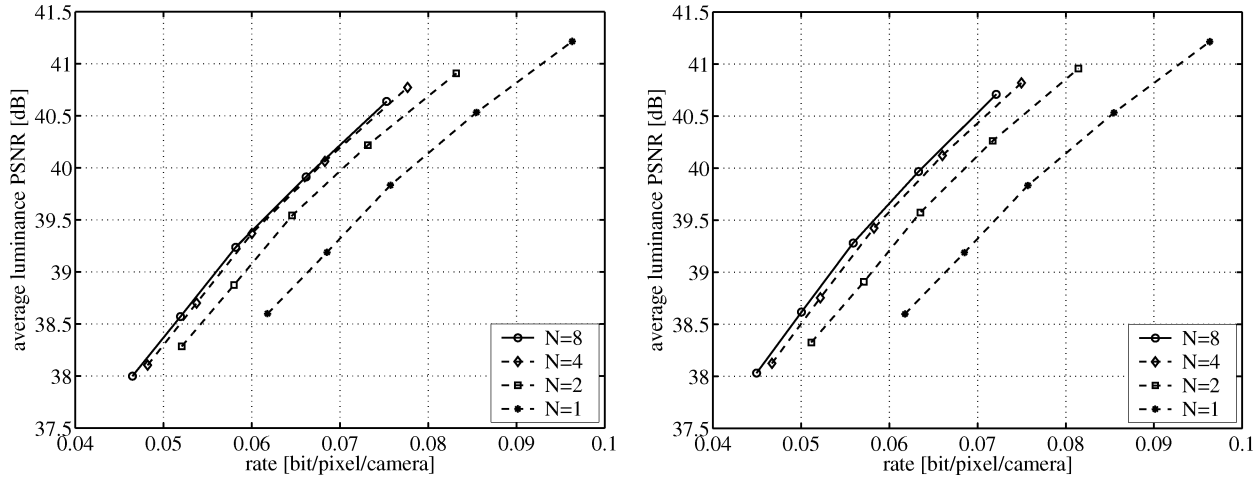

Fig. 5. Average luminance PSNR versus bit-rate for encoding 8 view-sequences of *Ballet* after histogram matching. The performance is plotted for a GOV size of $N = 1, 2, 4$, and 8. The disparity compensation is (left) integer-pel accurate and (right) quarter-pel accurate. The temporal GOP size is $K = 8$ and motion compensation is quarter-pel accurate.

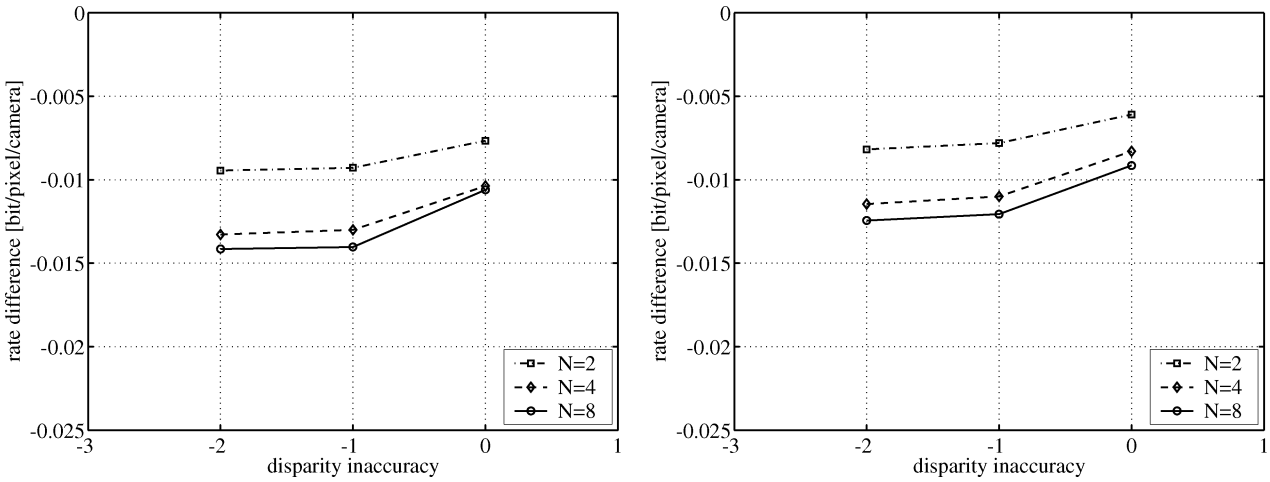

Fig. 6. Rate difference to independent encoding of each view sequence versus disparity inaccuracy of disparity compensation for 8 view-sequences of (left) histogram matched *Ballet* and (right) original *Ballet*. The performance is plotted for a GOV size of $N = 2$, 4, and 8, where $N = 1$ is the reference. The temporal GOP size is $K = 8$ and motion compensation is quarter-pel accurate. The rates are obtained for $\mathrm{PSNR} = 40$ dB.

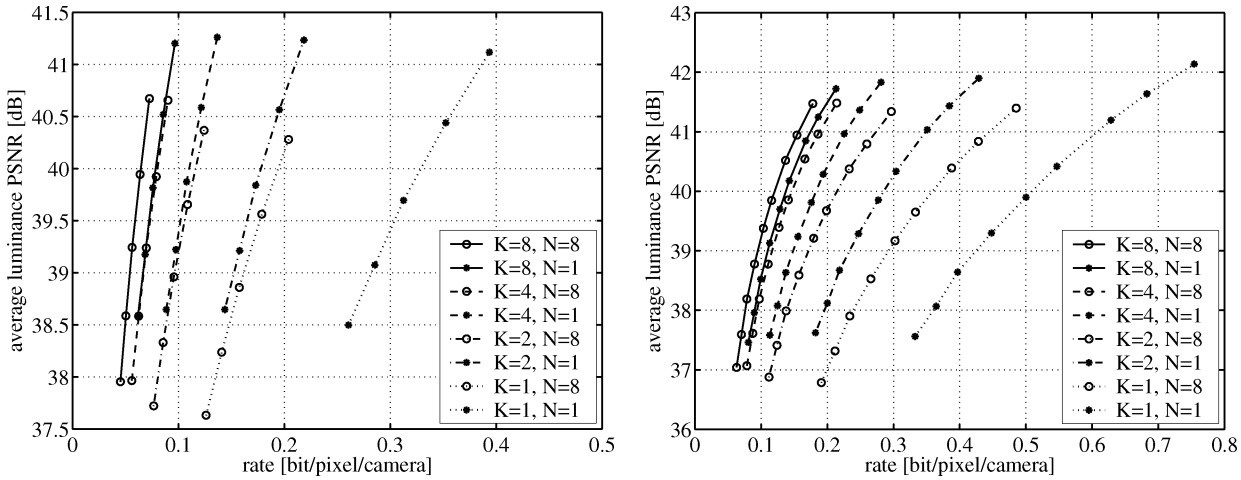

Fig. 7. Average luminance PSNR versus bit-rate for encoding eight view sequences of (left) histogram matched *Ballet* and of (right) *Exit*. The performance is plotted for a GOV size of $N = 1$ and 8 as well as a temporal GOP size of $K = 1$, 2, 4, and 8. Both disparity and motion compensation are quarter-pel accurate.

Gaussian noise $\mathbf{n}_k$. Fig. 11 summarizes the model. We use the same basic components to model both view and temporal correlation as our experimental codec uses the same coding technique for exploiting both of them. Note that all K temporal pictures of the νth view, $\nu = 1, 2, \ldots, N$, are shifted by the same disparity error vector $\mathbf{\Theta}_{1\nu}$, where the reference view is

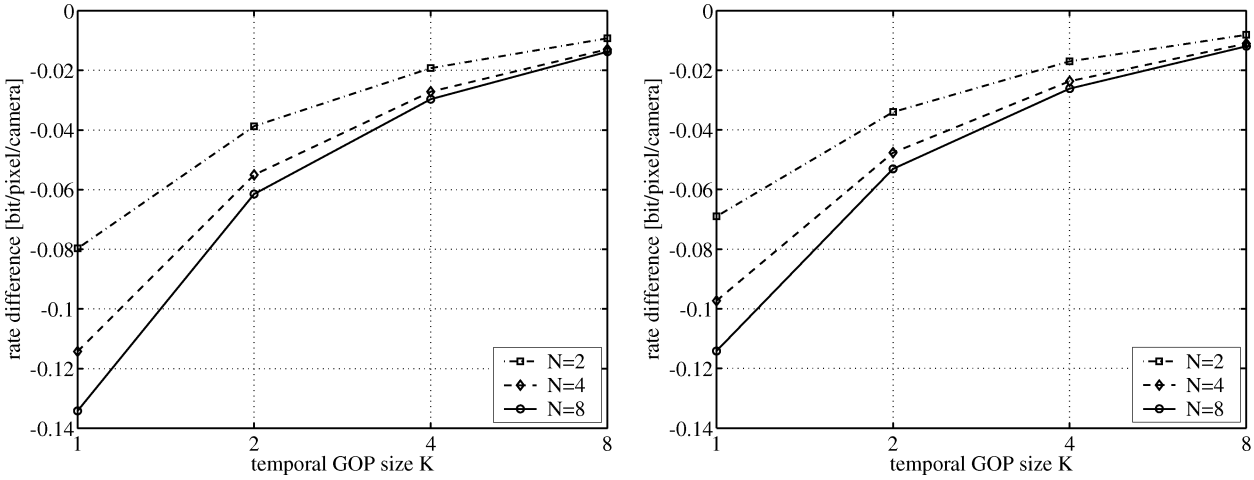

Fig. 8. Rate difference to independent encoding of each view sequence versus temporal GOP size K for eight view sequences of (left) histogram matched *Ballet* and (right) original *Ballet*. The performance is plotted for a GOV size of $N = 2, 4$, and 8, where $N = 1$ is the reference. Both disparity and motion compensation are quarter-pel accurate. The PSNR is 40 dB.

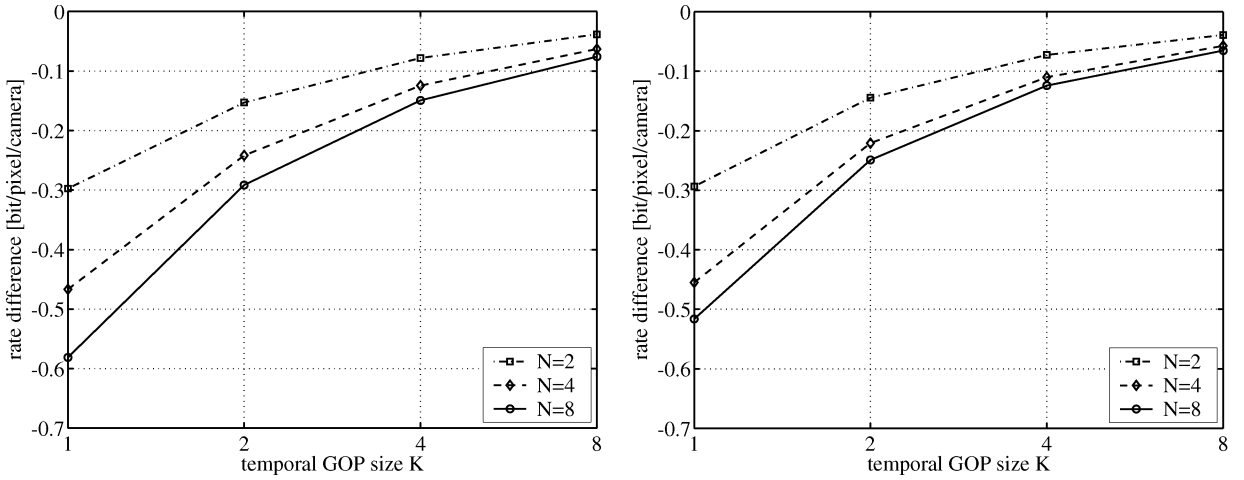

Fig. 9. Rate difference to independent encoding of each view sequence versus temporal GOP size K for eight view sequences of (left) histogram matched *Ballroom* and (right) *Race1*. The performance is plotted for a GOV size of $N = 2, 4$, and 8, where $N = 1$ is the reference. Both disparity and motion compensation are quarter-pel accurate. The PSNR is 40 dB.

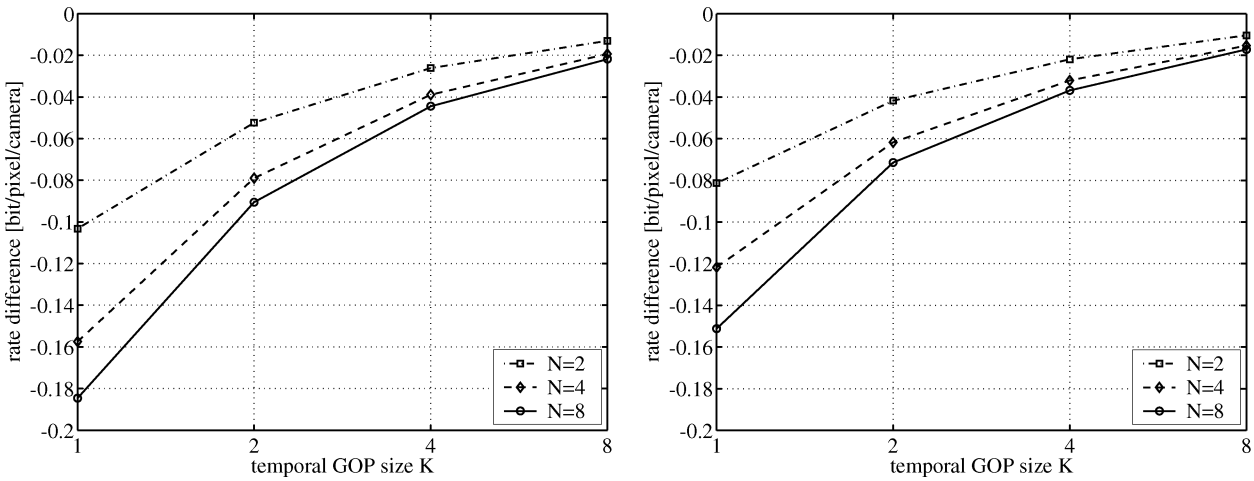

Fig. 10. Rate difference to independent encoding of each view sequence versus temporal GOP size K for eight view sequences of (left) *Breakdancers* and (right) *Exit*. The performance is plotted for a GOV size of $N = 2, 4$, and 8, where $N = 1$ is the reference. Both disparity and motion compensation are quarter-pel accurate. The PSNR is 40 dB.

the first view. We assume that the position of each camera is constant in time. Hence, we observe the same disparity error vector at each time instant. For example, consider the multiview video of a resting object. As the displacement errors are zero,

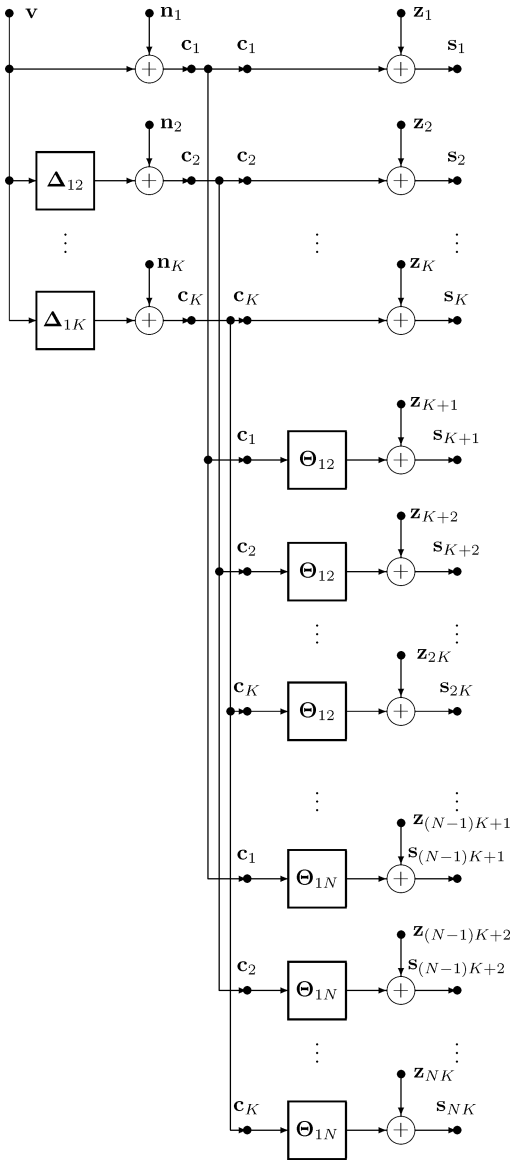

Fig. 11. Signal model for N image sequences each comprising of a group of K temporally successive pictures.

we observe the same disparity error vector at each time instant. For a moving object, displacement error vectors and disparity error vectors add up to form an individual shift error vector for each time instant of each view.

The work in [22] 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 $\boldsymbol{\Delta}_{\kappa\mu} + \boldsymbol{\Delta}_{\mu\nu} = \boldsymbol{\Delta}_{\kappa\nu}$. In the following, we assume also additive disparity, and consequently, additive disparity error $\boldsymbol{\Theta}_{\kappa\mu} + \boldsymbol{\Theta}_{\mu\nu} = \boldsymbol{\Theta}_{\kappa\nu}$. 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 assume that displacement errors and disparity errors are statistically independent.

Note that the model in [22] is included as the special case $N = 1$, i.e., only one view sequence. Moreover, the models are identical for $N = 1$ if the sums of the residual noise signals $\mathbf{n}_i + \mathbf{z}_i$ match the corresponding noise signals in [22].

Now, we adopt from [22] the PSD matrix of the root image sequence, normalized to the PSD of the root 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 video noise $\Phi_{\mathbf{n}_k\mathbf{n}_k}(\omega)$ with respect to the root picture $\mathbf{v}$

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

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

$$P(\omega) = E\left\{e^{-j\omega^T\boldsymbol{\Delta}_{\mu\nu}}\right\} = e^{-1/2\omega^T\omega\sigma^2_{\boldsymbol{\Delta}}}. \quad (3)$$

With the signal model in Fig. 11 and the previous assumptions for displacement and disparity errors, the PSD matrix of N view sequences each of length K 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 errors

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

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

$$G(\omega) = E\left\{e^{-j\omega^T\boldsymbol{\Theta}_{\mu\nu}}\right\} = e^{-1/2\omega^T\omega\sigma^2_{\boldsymbol{\Theta}}}. \quad (6)$$

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

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

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 root image sequence as we assume statistical independence between displacement errors and disparity errors.

B. Transform Coding Gain

Now, we use this mathematical model to determine rate-distortion bounds for multiview video signals. The practical coding scheme in Section II is a closed-loop predictive coder. But here, we are not interested in bounds for a particular coding scheme. We are rather interested in bounds for multiview imagery given

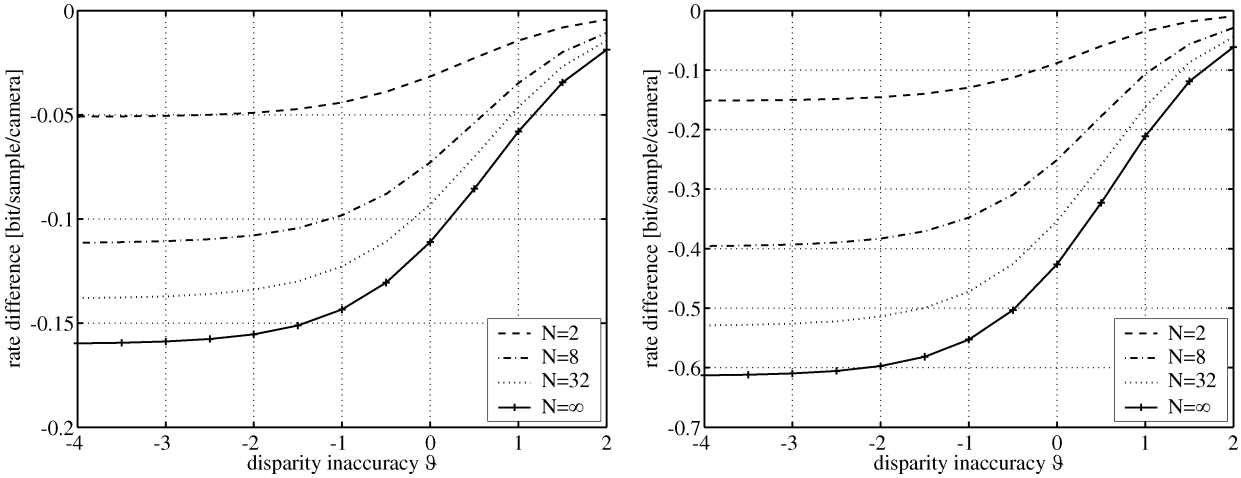

Fig. 12. Rate difference to independent encoding of each view sequence versus disparity inaccuracy ϑ of disparity compensation for GOV sizes of N. The displacement inaccuracy β of motion compensation among $K = 8$ pictures is -2 (quarter-pel accuracy) and the RMVNL is -10 dB. The RVNL is (left) -30 dB and (right) -10 dB.

parameters like the size of the MOP (N, K) or the accuracy of disparity compensation. This will help us to explain the experimental observations.

At high rates, rate-distortion bounds can be determined by assuming optimal transform coding with the KLT. For that, we calculate 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 $\Phi_{\mathbf{cc}}(\omega)/\Phi_{\mathbf{vv}}(\omega)$ 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 normalized eigenvalues of $\Phi_{\mathbf{ss}}(\omega)/\Phi_{\mathbf{vv}}(\omega)$ with their respective occurrences are

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

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:

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

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

We assess the performance of the multiview video coding scheme by using the average rate difference to independent encoding of N view sequences

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

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 [40].

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 video noise level* $\mathrm{RVNL} = 10\log_{10}(\sigma_{\mathbf{n}}^2)$ is -30 dB, which is typical for low-motion video sequences. We show both graphs also for a high residual video noise level of -10 dB. The *residual multiview noise level* $\mathrm{RMVNL} = 10\log_{10}(\sigma_{\mathbf{z}}^2)$ is -10 dB reflecting a large disparity model error to capture new scene content. Note that the root picture is normalized as $\sigma_{\mathbf{v}}^2 = 1$. The motion inaccuracy $\beta = \log_2(\sqrt{12}\sigma_{\mathbf{\Delta}})$ is a function of the variance of the displacement error components $\sigma_{\mathbf{\Delta}}^2$. 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. 12 depicts the average rate difference to independent encoding of each view sequence over the disparity inaccuracy ϑ of disparity compensation for a temporal GOP size of $K = 8$. The RVNL is (left) -30 dB and (right) -10 dB. To improve the readability of the graph, the disparity inaccuracy $\vartheta = \log_2(\sqrt{12}\sigma_{\mathbf{\Theta}})$ is a function of the variance of the disparity error components $\sigma_{\mathbf{\Theta}}^2$. 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 $N = 2$ views only. Experimental results in Figs. 4 and 6 show the same effect.

Fig. 13 depicts the rate difference in bit per sample per camera to independent encoding of each view sequence versus temporal GOP size K for various GOV sizes N. The RVNL is (left) -30 dB and (right) -10 dB. 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

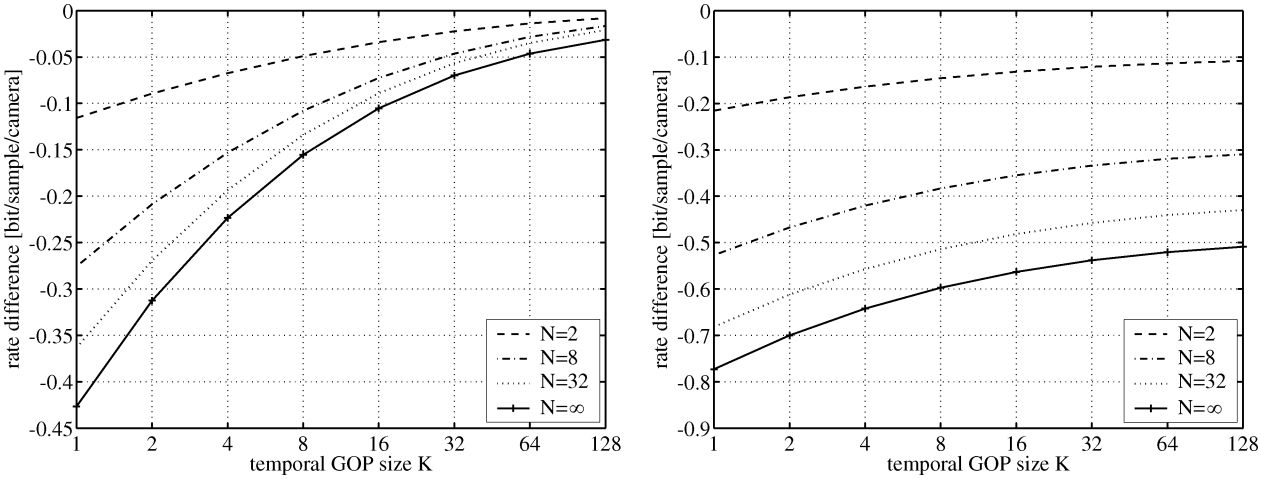

Fig. 13. Rate difference to independent encoding of each view sequence versus temporal GOP size K for groups of N views. The displacement inaccuracy β of motion compensation among K pictures as well as the disparity inaccuracy ϑ of disparity compensation among N views is -2 (quarter-pel accuracy). The RMVNL is -10 dB. The RVNL is (left) -30 dB and (right) -10 dB.

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 becomes weaker for smaller GOV size N. Experimental results in Figs. 8–10 show the same effect.

Note that there is a tradeoff when exploiting the correlation in temporal and view direction. If the residual video noise level is small, the reference coding scheme performs efficiently when exploiting temporal correlation. Hence, only a small margin is left for multiview video coding. On the other hand, if the residual video noise level is large, the reference coding scheme performs poorly when exploiting temporal correlation. Hence, a large margin is left for multiview video coding.

C. Special Cases

Finally, we are interested in the performance bound for very large MOP sizes, i.e., very large temporal GOP sizes K and GOV sizes N. In the limit, the eigenvalues $\lambda_2(\omega)$ and $\lambda_4(\omega)$ dominate the average rate difference

$$\Delta R_{N,K\to\infty} = \frac{1}{4\pi^2} \cdot \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{1}{2} \log_2 \frac{[1+\alpha(\omega)-P(\omega)]\,[1-G(\omega)]+\gamma(\omega)}{1+\alpha(\omega)-P(\omega)+\gamma(\omega)} d\omega. \quad (11)$$

Obviously, if we choose additionally very inaccurate disparity compensation, i.e., $\vartheta \to \infty$, and hence, $G(\omega) \to 0$, the average rate difference approaches zero

$$\Delta R_{N,K\to\infty;G\to 0} = 0. \quad (12)$$

This is reasonable as very inaccurate disparity compensation is not able to exploit the view correlation. A joint coding scheme with such an inaccurate disparity compensation will not provide any benefit. On the other hand, very accurate disparity compensation, i.e., $\vartheta \to -\infty$, and hence, $G(\omega) \to 1$, will offer rate savings according to

$$\Delta R_{N,K\to\infty;G\to 1} = \frac{1}{4\pi^2} \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{1}{2} \log_2 \frac{\gamma(\omega)}{1+\alpha(\omega)-P(\omega)+\gamma(\omega)} d\omega. \quad (13)$$

With very accurate motion and disparity compensation, i.e., $P(\omega) \to 1$ and $G(\omega) \to 1$, the performance bound for very large MOP sizes simplifies to

$$\Delta R_{N,K\to\infty;P,G\to 1} = \frac{1}{4\pi^2} \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{1}{2} \log_2 \frac{\gamma(\omega)}{\alpha(\omega)+\gamma(\omega)} d\omega. \quad (14)$$

Thus, the average rate difference depends on the relative strength of video noise and multiview noise. In case the normalized power spectral density of the video noise is equal to that of the multiview noise, i.e., $\alpha(\omega) = \gamma(\omega)$, the average rate difference to independent encoding is -0.5 bit per sample per camera. This can be seen in the right plot of Fig. 13.

If the multiview noise is substantially larger than the video noise, joint encoding provides only a limited advantage over independent encoding. On the other hand, if the multiview noise is substantially smaller than the video noise, joint encoding offers a substantial benefit over independent encoding of the view sequences.

IV. Conclusion

We experimentally and theoretically study the problem of coding N multiview video sequences. We define a matrix of pictures with N view sequences, each with K temporally successive pictures. We devise a coding scheme based on H.264/AVC and utilize histogram matching to compensate for inter-view intensity variations. For groups of N views (GOV), we discuss the impact of both inaccurate disparity compensation and temporal GOP size K on the overall rate-distortion efficiency. We observe

that the efficiency improves with accurate disparity compensation. Moreover, the relative efficiency grows with increasing GOV size N. But with increasing temporal GOP size K, the relative efficiency grows slower with GOV size N. Finally, we propose and discuss a high-rate model for multiview video coding that explains our experimental observations.

ACKNOWLEDGMENT

The authors would like to thank Mitsubishi Electric Research Laboratories, KDDI R&D Laboratories, the Interactive Visual Media Group at Microsoft Research, and Tanimoto Laboratory at Nagoya University who kindly provided multiview video test sequences.

REFERENCES

[1] A. Smolic and P. Kauff, "Interactive 3-D video representation and coding technologies," *Proc. IEEE*, vol. 93, no. 1, pp. 98–110, Jan. 2005.
[2] N. A. Dodgson, "Autostereoscopic 3-D displays," *IEEE Comput.*, vol. 38, no. 8, pp. 31–36, Aug. 2005.
[3] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 765–776, Jul. 2005.
[4] M. Tanimoto, "Free viewpoint television – FTV," presented at the Picture Cod. Symp., San Francisco, CA, 2004.
[5] M. Tanimoto, "FTV (free viewpoint television) creating ray-based image engineering," presented at the IEEE Int. Conf. Image Process., Genova, Italy, 2005.
[6] M. Tanimoto, "FTV (free viewpoint television) for 3-D scene reproduction and creation," presented at the IEEE Conf. Comput. Vision Pattern Recog., New York, NY, 2006.
[7] O. Schreer, C. Fehn, N. Atzpadin, M. Muller, A. Smolic, R. Tanger, and P. Kauff, "A flexible 3-D TV system for different multi-baseline geometries," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2006, pp. 1877–1880.
[8] A. Smolic, K. Mueller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, and T. Wiegand, "3-D video and free viewpoint video – Technologies, applications and MPEG standards," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2006, pp. 2161–2164.
[9] R. S. Wang and Y. Wang, "Multiview video sequence analysis, compression, and virtual viewpoint synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 4, pp. 397–410, Apr. 2000.
[10] K. Mueller, P. Merkle, H. Schwarz, T. Hinz, A. Smolic, T. Oelbaum, and T. Wiegand, "Multi-view video coding based on H.264/AVC using hierarchical B-frames," presented at the Picture Cod. Symp., Beijing, China, 2006.
[11] E. Martinian, A. Behrens, J. Xin, and A. Vetro, "View synthesis for multiview video compression," presented at the Picture Cod. Symp., Beijing, China, 2006.
[12] K. Yamamoto, T. Yendo, T. Fujii, M. Tanimoto, M. Kitahara, H. Kimata, S. Shimizu, K. Kamikura, and Y. Yashima, "Multi-view video coding using view-interpolated reference images," presented at the Picture Cod. Symp., Beijing, China, 2006.
[13] S.-C. Chan, K.-T. Ng, Z.-F. Gan, K.-L. Chan, and H.-Y. Shum, "The compression of simplified dynamic light fields," presented at the IEEE Int. Conf. Acoust., Speech Signal Process., Hong Kong, 2003.
[14] Z.-F. Gan, S.-C. Chan, K.-T. Ng, K.-L. Chan, and H.-Y. Shum, "On the rendering and post-processing of simplified dynamic light fields with depth information," presented at the IEEE Int. Conf. Acoust., Speech Signal Process., Montreal, QC, Canada, 2004.
[15] B. Girod, C.-L. Chang, P. Ramanathan, and X. Zhu, "Light field compression using disparity-compensated lifting," presented at the IEEE Int. Conf. Acoust., Speech Signal Process., Hong Kong, 2003.
[16] C.-L. Chang, X. Zhu, P. Ramanathan, and B. Girod, "Light field compression using disparity-compensated lifting and shape adaptation," *IEEE Trans. Image Process.*, vol. 15, no. 4, pp. 793–806, Apr. 2006.
[17] W. Yang, F. Wu, Y. Lu, J. Cai, K. N. Ngan, and S. Li, "Scalable multiview video coding using wavelet," presented at the IEEE Int. Symp. Circuits Syst., Kobe, Japan, 2005.
[18] W. Yang, F. Wu, Y. Lu, J. Cai, K. N. Ngan, and S. Li, "4-D wavelet-based multiview video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 11, pp. 1385–1396, Nov. 2006.
[19] B. Girod, "The efficiency of motion-compensating prediction for hybrid coding of video sequences," *IEEE J. Select. Areas Commun.*, vol. SAC-5, no. 7, pp. 1140–1154, Aug. 1987.
[20] M. Flierl and B. Girod, "Video coding with motion compensation for groups of pictures," in *Proc. IEEE Int. Conf. Image Process.*, 2002, pp. 69–72.
[21] M. Flierl and B. Girod, "Investigation of motion-compensated lifted wavelet transforms," in *Proc. Picture Cod. Symp.*, 2003, pp. 59–62.
[22] M. Flierl and B. Girod, "Video coding with motion-compensated lifted wavelet transforms," *Signal Process.: Image Commun.*, vol. 19, no. 7, pp. 561–575, Aug. 2004.
[23] A. Vetro, W. Matusik, H. Pfister, and J. Xin, "Coding approaches for end-to-end 3-D TV systems," presented at the Picture Cod. Symp., San Francisco, CA, 2004.
[24] M. Flierl, A. Mavlankar, and B. Girod, "Motion and disparity compensated coding for video camera arrays," presented at the Picture Cod. Symp., Beijing, China, 2006.
[25] P. Ramanathan and B. Girod, "Theoretical analysis of geometry inaccuracy for light field compression," in *Proc. IEEE Int. Conf. Image Process.*, 2002, pp. 229–232.
[26] P. Ramanathan and B. Girod, "Rate-distortion analysis of random access for compressed light fields," in *Proc. IEEE Int. Conf. Image Process.*, 2004, pp. 2463–2466.
[27] P. Ramanathan and B. Girod, "Theoretical analysis of the rate-distortion performance of a light field streaming system," presented at the Picture Cod. Symp., San Francisco, CA, 2004.
[28] P. Ramanathan, "Compression and interactive streaming of light fields," Ph.D. dissertation, Dept. Electr. Eng., Stanford University, Stanford, CA, 2005.
[29] P. Ramanathan and B. Girod, "Rate-distortion analysis for light field coding and streaming," *Signal Process.: Image Commun.*, vol. 21, no. 6, pp. 462–475, Jul. 2006.
[30] A. Smolic and D. McCutchen, "3DAV exploration of video-based rendering technology in MPEG," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 3, pp. 348–356, Mar. 2004.
[31] A. Vetro, Y. Su, H. Kimata, and A. Smolic, "Joint multiview video model JMVM 2.0," ITU-T and ISO/IEC Joint Video Team, Doc. JVT-U207, 2006.
[32] "ITU-T Rec. H.264 – ISO/IEC 14496–10 AVC: Advanced video coding for generic audiovisual services," ITU-T and ISO/IEC Joint Video Team, 2005.
[33] M. Flierl and B. Girod, "Generalized B pictures and the draft H.264/AVC video compression standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 587–597, Jul. 2003.
[34] M. Flierl and B. Girod, "Multihypothesis motion estimation for video coding," in *Proc. Data Compress. Conf.*, 2001, pp. 341–350.
[35] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.
[36] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1995.
[37] U. Fecker, M. Barkowsky, and A. Kaup, "Improving the prediction efficiency for multi-view video coding using histogram matching," presented at the Picture Cod. Symp., Beijing, China, 2006.
[38] A. Kaup and U. Fecker, "Analysis of multi-reference block matching for multi-view video coding," presented at the Workshop Dig. Broadcast., Erlangen, Germany, 2006.
[39] W. Li, J.-R. Ohm, M. van der Schaar, H. Jiang, and S. Li, "MPEG-4 video verification model version 18.0," ISO/IEC JTC1/SC29/WG11, Doc. MPEG N3908, 2001.
[40] B. Girod, "Efficiency analysis of multihypothesis motion-compensated prediction for video coding," *IEEE Trans. Image Process.*, vol. 9, no. 2, pp. 173–183, Feb. 2000.

Markus Flierl (S'01–M'04) received the Ph.D. degree in engineering from Friedrich Alexander University, Erlangen, Germany, in 2003.

He is currently a Visiting Assistant Professor with the Max Planck Center for Visual Computing and Communication, Stanford University, Stanford, CA. From 2000 to 2002, he visited the Information Systems Laboratory, Stanford University. From 2003 to 2005, he was a Senior Researcher with the Signal Processing Institute, Swiss Federal Institute of Technology, Lausanne, Switzerland. He is the author of the book *Video Coding with Superimposed Motion-Compensated Signals: Applications to H.264 and Beyond*. His research interests include visual communication networks and video representations.

Dr. Flierl was the recipient of the VCIP 2007 Young Investigator Award.

Aditya Mavlankar (S'99) received the B.E. degree in electronics and telecommunication from University of Pune, Pune, India, in 2002 and the M.S. degree in communications engineering from the Technical University of Munich, Munich, Germany, in 2004.

He is currently a Researcher with the Information Systems Laboratory, Stanford University, Stanford, CA. His research interests include scalable video coding, interactive video delivery and peer-to-peer video streaming.

Mr. Mavlankar was the recipient of the Edison Prize Bronze Medal by IIE Europe in conjunction with the GE Foundation for his Master's thesis in 2006, a corecipient of the Best Student Paper Award at the IEEE Workshop on Multimedia Signal Processing (MMSP), Victoria, BC, Canada.

Bernd Girod (M'80–SM'97–F'98) received the Ph.D. degree in engineering from University of Hannover, Hannover, Germany, and the M.S. degree from Georgia Institute of Technology, Atlanta.

He is a Professor of electrical engineering and (by courtesy) computer science with the Information Systems Laboratory, Stanford University, Stanford, CA. He was Chaired Professor of Telecommunications in the Electrical Engineering Department, the University of Erlangen-Nuremberg, Erlangen, Germany, from 1993 to 1999. His research interests include the areas of video compression and networked media systems. Prior visiting or regular faculty positions include MIT, Georgia Institute of Technology and Stanford University. He has been involved with several startup ventures as founder, director, investor, or advisor, among them Vivo Software, 8×8 (Nasdaq: EGHT) and RealNetworks (Nasdaq: RNWK). Since 2004, he has served as the Chairman of the new Deutsche Telekom Laboratories, Berlin, Germany.