Available online at www.sciencedirect.com
SCIENCE DIRECT
Signal Processing: *Image Communication* 21 (2006) 462–475
SIGNAL PROCESSING: IMAGE COMMUNICATION
www.elsevier.com/locate/image

Rate-distortion analysis for light field coding and streaming

Prashant Ramanathan*,[1], Bernd Girod

Department of Electrical Engineering, Stanford University, USA

Received 1 March 2006; accepted 2 March 2006

Abstract

A theoretical framework to analyze the rate-distortion performance of a light field coding and streaming system is proposed. This framework takes into account the statistical properties of the light field images, the accuracy of the geometry information used in disparity compensation, and the prediction dependency structure or transform used to exploit correlation among views. Using this framework, the effect that various parameters have on compression efficiency is studied. The framework reveals that the efficiency gains from more accurate geometry, increase as correlation between images increases. The coding gains due to prediction suggested by the framework match those observed from experimental results. This framework is also used to study the performance of light field streaming by deriving a view-trajectory-dependent rate-distortion function. Simulation results show that the streaming results depend both the prediction structure and the viewing trajectory. For instance, independent coding of images gives the best streaming performance for certain view trajectories. These and other trends described by the simulation results agree qualitatively with actual experimental streaming results.

Keywords: Light fields; Light field coding; Light field streaming; Rate-distortion theory; Statistical signal processing

1. Introduction

From fly-arounds of automobiles to 360° panoramic views of cities to walk-throughs of houses, 3-D content is increasingly becoming commonplace on the Internet. Current content, however, offers limited mobility around the object or scene, and limited image resolution. To generate high-quality photo-realistic renderings of 3-D objects and scenes, computer graphics has traditionally turned to computationally expensive algorithms such as ray-tracing.

Recently, there has been increasing attention on image-based rendering techniques that require only resampling of captured images to render novel views. Such approaches are especially useful for interactive applications because of their low rendering complexity. A *light field* [24,19] is an image-based rendering dataset that allows for photo-realistic rendering quality, as well as much greater freedom in navigating around the scene or object. To achieve this, light fields rely on a large number of captured images.

The large amount of data can make transmitting light fields from a central server to a remote user a challenging problem. The sizes of certain large datasets can be in the tens of Gigabytes [25]. Even over fast network connections, it could take hours

*Corresponding author.
E-mail address: pramanat@gmail.com (P. Ramanathan).
[1]Now with NetEnrich, Inc., Santa Clara, CA, USA.

0923-5965/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2006.03.002

to download the raw data for a large light field. This motivates the need for efficient compression algorithms to reduce the data size.

Numerous algorithms have been proposed to compress light fields. The most efficient algorithms use a technique called disparity compensation. Disparity compensation uses either implicit geometry information or an explicit geometry model to warp one image to another image. This allows for prediction between images, such as in [28,42,34,26,29,6,39], encoding several images jointly by warping them to a common reference frame, as in [31,27,29,40], or, more recently, by combining prediction with lifting, as in [5,18,6]. This paper considers a closed-loop predictive light field coder [38,6] from the first set of techniques.

In this light field coder, an explicit geometry model is used for disparity compensation. For real-world sequences, this geometry model is estimated from image data using computer vision techniques, which introduces inaccuracies. In addition, lossy encoding is used to compress the description of the geometry model, which leads to additional inaccuracies in the geometry data. Light field images, captured in a hemispherical camera arrangement, are organized into different levels. The first level is a set of key images, where each image is independently compressed without predicting from nearby images. These images are evenly distributed around the hemisphere. Each image in the second level uses the two nearest neighboring images in the first level to form a prediction image with disparity compensation. The residual image resulting from prediction is encoded. The images in the third levels uses the nearest images in the first two levels for prediction, and so on, until all the images in the light field are encoded.

As theoretical and experimental results later in this paper will show, using prediction typically improves compression efficiency. It also, however, creates decoding dependencies between images. While this is acceptable for a download scenario where the entire light field is retrieved and decoded, for other transmission scenarios such as interactive streaming, decoding dependencies may affect the performance of the system. This paper considers an interactive light field streaming scenario where a user remotely accesses a dataset over a best-effort packet network. As the user starts viewing the light field dataset by navigating around the scene or object, the encoded images that are appropriate for rendering the desired view are transmitted [37,35].

For interactive performance, there are stringent latency constraints on the system. This means that a view, once selected by the user, must be rendered by a particular deadline. Due to the best-effort nature of the network and possible bandwidth limitations, not all images required for rendering that view may be available by the rendering deadline. In this case, the view is rendered with a subset of the available set of images, resulting in degraded image quality. The decoding dependencies due to prediction between images will affect the set of images that can be decoded, and thus the rendered image quality and the overall streaming performance.

This paper presents a theoretical framework to study the rate-distortion performance of a light field coding and streaming system. Coding and streaming performance is affected by numerous factors including geometry accuracy, the prediction dependency structure, between-view and within-view correlation, the image selection criteria for rendering, the images that are successively transmitted and the user's viewing trajectory. The framework that is presented incorporates these factors into a tractable model.

This work is based on the theoretical analysis of the relationship between motion compensation accuracy and video compression efficiency [15–17,8,9]. It is also closely related to the theoretical analysis of 3-D subband coding of video [10–12]. Prior theoretical work has analyzed light fields from a sampling perspective [4,21,41], but not a rate-distortion perspective as in this work.

This paper is organized as follows. The theoretical framework for coding the entire light field is described in Section 2, along with simulation and experimental coding results. Section 3 presents the extension to light field streaming, along simulation results and comparison to the results of light field streaming experiments.

2. Compression performance for the entire light field

2.1. Geometric model

The formation of light field images depends on the complex interaction of lighting with the scene geometry, as well as surface properties, camera parameters and imaging noise. A simple model of how light field images are generated is required for a tractable rate-distortion analysis of light field coding. This simplification starts by modeling the complex 3-D object or scene represented by the light field as a planar surface. On this planar surface is a

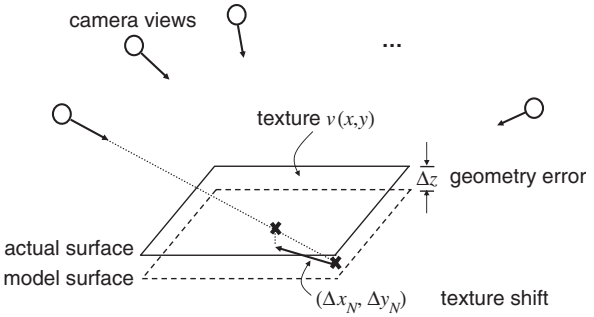

Fig. 1. Light field images of a planar object.

2-D texture signal $v(x,y)$ that is viewed by N cameras from directions $r_1, r_2, \ldots, r_N$. This arrangement is illustrated in Fig. 1.

The light field is captured with parallel projection cameras. This assumption allows the camera to be parameterized in terms of the camera direction, without need of knowing the exact camera position. This simplifies the derivation when later modeling geometry inaccuracy. The camera is also assumed to suffer no bandlimitation restriction due to imaging resolution limits, as would a real camera viewing the plane at a grazing angle.

The planar surface that represents the object is considered to be approximately Lambertian, that is, its appearance is similar from all viewpoints. Any non-Lambertian or view-dependent effects are modeled as additive noise. Included in this noise term is any image noise from the camera.

When predicting or warping one image to one another, it is useful to consider the images in the texture domain. A camera image can be back-projected onto the planar geometry to generate a texture image that is very similar to the original texture v. The geometry, however, is not accurately known. The geometry error can be modeled as an offset Δz of the planar surface from its true position, as illustrated in Fig. 1.

When the camera image is back-projected from view i onto this inaccurate planar surface, this results in a texture image c_i that is a shifted version of the original texture signal v. An additive noise term n_i accounts for effects such as image noise, non-Lambertian view-dependent effects, occlusion or any aspect of geometry compensation that is not described by this simple model of geometry compensation. The back-projected texture image is given by the equation

$$c_i(x,y) = v(x - \Delta_{xi}, y - \Delta_{yi}) + n_i(x,y). \tag{1}$$

The shift, which depends only upon the camera's viewing direction $\mathbf{r}_i = [r_{ix}\ r_{iy}\ r_{iz}]^{\mathrm{T}}$ and the geometry error Δz, is described by the equation

$$\begin{bmatrix} \Delta_x \\ \Delta_y \end{bmatrix} = \frac{\Delta z}{r_{iz}} \begin{bmatrix} r_{ix} \\ r_{iy} \end{bmatrix}. \tag{2}$$

As the eventual goal is a frequency-domain analysis of these signals, the transfer function of the shift can be represented by $S_i(\omega_x, \omega_y) = \mathrm{e}^{-\mathrm{j}(\omega_x \Delta_{xi} + \omega_y \Delta_{yi})}$.

The image vector $\mathbf{c} = [c_1\ c_2\ \cdots\ c_N]^{\mathrm{T}}$ represents the set of light field images or texture maps that have already been compensated or corrected with the true geometry, up to the inaccuracy (Δ_x, Δ_y). A light field coder does not encode these geometry-compensated images independently, but first tries to exploit the correlation between them. Note that perfect knowledge of the geometry would mean that the geometry-compensated images are perfectly aligned.

Prediction of light field images in the texture domain is simply a matter of subtracting one image from another or, more generally, taking a linear combination of images. Wavelet or subband coding across images can similarly be described with a linear combination of images. Taken across all images, a linear transform T describes either a prediction-based or transform-based scheme, that attempts to remove the correlation between the geometry-compensated light field images $\mathbf{c} = [c_1\ c_2\ \cdots\ c_N]^{\mathrm{T}}$. The result of this transform, as shown in Fig. 2, is a set of residual error images, or coefficient images, $\mathbf{e} = [e_1\ e_2\ \cdots\ e_N]^{\mathrm{T}}$. Each of these error images is finally independently coded.

Strictly speaking, only an open-loop transform across the images is correctly modeled by use of the transform T in the model. In closed-loop prediction, images are predicted from reconstructed images and

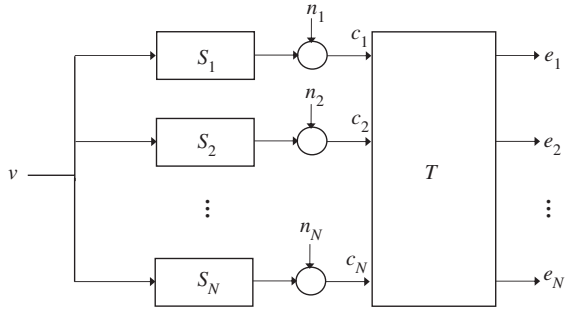

Fig. 2. Signal model for generation and coding of light field images.

not the original images. Previous analyses of DPCM systems such as in [7,22] have assumed that the effect of quantization errors on prediction efficiency is negligible for sufficiently fine quantization. The observations from these analyses agree reasonably well with experiments. In the following analysis, the assumption that the effect of quantization errors on prediction efficiency can be neglected for sufficiently fine quantization is also made.

2.2. Statistical model

The texture signal v is modeled as a wide-sense stationary random process, that has been appropriately bandlimited and sampled on a grid with unit spacing in the x and y directions. The power spectral density (PSD) $\Phi_{vv}(\omega_x, \omega_y)$ of this signal is defined over $\omega_x \in [-\pi, \pi]$, $\omega_y \in [-\pi, \pi]$. The PSD of a signal can be derived as the discrete Fourier transform of its autocorrelation function.

The noise images $\mathbf{n} = n_0, n_1, \ldots, n_N$ are also defined in a similar manner, as jointly wide-sense stationary signals. The cross-correlation between n_i and n_j is given by the PSD $\Phi_{n_i n_j}$, which can be collected into an overall PSD matrix $\mathbf{\Phi}_{nn}$ for $\mathbf{n}$. The noise term $\mathbf{n}$ is assumed to be independent of the texture signal v.

By collecting the set of shift transfer functions $\{S_i\}$ into a column vector

$$S = \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_N \end{bmatrix} = \begin{bmatrix} \mathrm{e}^{-\mathrm{j}(\omega_x \Delta_{x1} + \omega_y \Delta_{y1})} \\ \mathrm{e}^{-\mathrm{j}(\omega_x \Delta_{x2} + \omega_y \Delta_{y2})} \\ \vdots \\ \mathrm{e}^{-\mathrm{j}(\omega_x \Delta_{xN} + \omega_y \Delta_{yN})} \end{bmatrix}, \tag{3}$$

the PSD of corresponding vector signal $\mathbf{c}$ is

$$\mathbf{\Phi}_{cc} = \Phi_{vv} S S^{\mathrm{H}} + \mathbf{\Phi}_{nn}, \tag{4}$$

where the superscript H represents the complex-conjugate transpose of a matrix. Since $\mathbf{c}$ and $\mathbf{n}$ are vectors of size N, both $\mathbf{\Phi}_{cc}$ and $\mathbf{\Phi}_{nn}$ are matrices of size $N \times N$. The independent variables (ω_x, ω_y) have been omitted, but each term in (4) depends on them.

In (4) the geometry error Δz which influences the vector S is a deterministic quantity. Letting Δz be a random variable, independent of $\mathbf{c}$ and $\mathbf{n}$, with a probability density function (pdf) p, the revised equation,

$$\mathbf{\Phi}_{cc} = \Phi_{vv} E\{S S^{\mathrm{H}}\} + \mathbf{\Phi}_{nn}, \tag{5}$$

is obtained, where

$$E\{S S^{\mathrm{H}}\} = \begin{bmatrix} 1 & \ldots & P(\Omega^{\mathrm{T}}(\mathbf{a}_1 - \mathbf{a}_N)) \\ P(\Omega^{\mathrm{T}}(\mathbf{a}_2 - \mathbf{a}_1)) & \ldots & P(\Omega^{\mathrm{T}}(\mathbf{a}_2 - \mathbf{a}_N)) \\ \vdots & \ddots & \vdots \\ P(\Omega^{\mathrm{T}}(\mathbf{a}_N - \mathbf{a}_1)) & \ldots & 1 \end{bmatrix}. \tag{6}$$

$P(\omega = \Omega^{\mathrm{T}}(\mathbf{a}_i - \mathbf{a}_j))$ represents the 1-D Fourier transform of the pdf p of the random variable Δz, $\Omega = [\omega_x \ \omega_y]^{\mathrm{T}}$ is a vector quantity, and $\mathbf{a}_i = [r_{ix} \ r_{iy}]^{\mathrm{T}}/r_{iz}$ depends on the original viewing direction vectors $\mathbf{r}_i$.

The power spectrum of the error image signal $\mathbf{e}$ is

$$\begin{aligned} \mathbf{\Phi}_{ee} &= T \mathbf{\Phi}_{cc} T^{\mathrm{H}} \\ &= \Phi_{vv} T E\{S S^{\mathrm{H}}\} T^{\mathrm{H}} + T \mathbf{\Phi}_{nn} T^{\mathrm{H}}, \end{aligned} \tag{7}$$

where T is the linear transformation matrix described in the previous section. Again, note that $\mathbf{\Phi}_{ee}$ is a matrix of size $N \times N$.

2.3. Rate-distortion performance

Knowing the PSD of the residual error images (7), it is possible to derive the rate-distortion performance for each of these images. The rate-distortion function for a 2-D stationary Gaussian random process x with PSD Φ_{xx} is given in parametric form, with rate

$$R(\phi) = \frac{1}{8\pi^2} \int_{\omega_x=-\pi}^{\pi} \int_{\omega_y=-\pi}^{\pi} \max\left[0, \log_2 \frac{\Phi_{xx}(\omega_x, \omega_y)}{\phi}\right] \mathrm{d}\omega_x \, \mathrm{d}\omega_y \tag{8}$$

and distortion

$$D(\phi) = \frac{1}{4\pi^2} \int_{\omega_x=-\pi}^{\pi} \int_{\omega_y=-\pi}^{\pi} \min[\phi, \Phi_{xx}(\omega_x, \omega_y)] \, \mathrm{d}\omega_x \, \mathrm{d}\omega_y, \tag{9}$$

where the parameter $\phi \in [0, +\infty)$ traces out the rate-distortion curve [23,32,13,2].

Thus, if the signals v and $\mathbf{n}$ are assumed to be jointly Gaussian and stationary, implying $\mathbf{c}$ and $\mathbf{e}$ are also stationary and Gaussian, then the rate-distortion function for $\mathbf{c}$ and $\mathbf{e}$ can be determined by (8) and (9). This does not necessarily hold for a closed-loop system. Nevertheless, under the assumptions of high-rate uniform quantization,

the energy of the quantization error asymptotically tends to zero, thus the effects of quantization errors on prediction efficiency may be neglected. With this model, the coding gain of using either prediction or transform coding across images can be calculated.

In [16,17,10], a related performance measure, the rate difference, is used to measure this coding gain for high rates. The rate difference, at high rates, of coding the signal e_i instead of the signal c_i is given by

$$\begin{aligned}\Delta R_i &= R_{e_i}(\phi) - R_{c_i}(\phi) \\ &= \frac{1}{8\pi^2}\int_{\omega_x=-\pi}^{\pi}\int_{\omega_y=-\pi}^{\pi} \log_2 \frac{\Phi_{e_i e_i}(\omega_x,\omega_y)}{\Phi_{c_i c_i}(\omega_x,\omega_y)}\,\mathrm{d}\omega_x\,\mathrm{d}\omega_y\end{aligned} \tag{10}$$

which can be found by substituting (8) into the first line of (10). Only the difference in rate needs to be considered, since the coding distortion for both signals is approximately identical at high rates. In the rest of the analysis, although an assumption of high rate is used, the rate-distortion function, as given in (8) and (9), will be the main focus.

The rate-distortion function for the residual error signals needs to be related to the rate-distortion performance for the entire light field. In a light field coder, the error residual images $\{e_i\}$ in Fig. 2 are reconstructed, giving reconstructed error residual images $\{\hat{e}_i\}$, and then inverse transformed to produce the reconstructed images $\{\hat{c}_i\}$. The distortion of a light field, due to coding, is measured between the original and reconstructed images as

$$\begin{aligned}D(\phi) &= \frac{1}{N}\sum_{i=1}^{N} D_i(\phi) \\ &= \frac{1}{N}\sum_{i=1}^{N} E\{(\hat{c}_i - c_i)^2\}.\end{aligned}$$

How the distortion term is related to the coding distortion of the error residual images, depends on whether the transform T refers to closed-loop prediction or open-loop transform across images. For closed-loop prediction, the quantization distortion between the reconstructed and original light field images is identical to that between the reconstructed and original residual images, as discussed, for instance, in [14]:

$$\hat{c}_i - c_i = \hat{e}_i - e_i. \tag{11}$$

Thus, the light field distortion due to coding for the closed-loop case can be written as

$$D(\phi) = \frac{1}{N}\sum_{i=1}^{N} E\{(\hat{e}_i - e_i)^2\} \tag{12}$$

$$= \frac{1}{N}\sum_{i=1}^{N} D_{e_i}(\phi), \tag{13}$$

where $D_{e_i}(\phi)$ is defined in (9), substituting e_i for x.

For the open-loop case, it is assumed that the transform T is unitary. Then, it can be shown [33] that the distortion can be written as $D(\phi) = (1/N)\sum_{i=1}^{N} D_{e_i}(\phi)$, identical to the closed-loop case (13). Therefore, for both the open-loop and closed-loop cases, the overall distortion of the light field is written as the sum of the distortion for each residual error image. The operational rate for independently encoding the residual error images, on the other hand, is simply measured as the sum of rates,

$$R(\phi) = \sum_{i=1}^{N} R_{e_i}(\phi), \tag{14}$$

where $R_{e_i}(\phi)$ is defined in (8), substituting e_i for x. In (13) and (14), an identical value of ϕ is applied to all images. In general, a different value ϕ_i could be used for each image i. However, at high rates, $D_{e_i}(\phi_i) \approx \phi_i$ and, typically, constant quality or distortion is desired for each image. Thus, $\phi_i = \phi$ can be used for all images i. Note that, in this analysis, the bit-rate for the geometry is neglected.

2.4. Simulation results

According to the theoretical model, the compression efficiency of light field encoding is affected by several different factors: the spatial correlation ρ of the image; the view-dependent image noise variance σ_N^2, which is a measure of the correlation between images; the geometry error variance σ_G^2; the camera viewing directions $\{\mathbf{r}_i\}$; and the prediction dependency structure, captured by the matrix T.

The geometry error z is modeled as a zero-mean Gaussian random variable with variance σ_G^2. In light of the simplifying assumption of a planar geometry, the exact shape of the pdf is not important. Rather, the salient point is that by varying σ_G^2 the effect of geometry error can be studied. An isotropic, exponentially decaying autocorrelation function in the form

$$R_{vv}(\tau_x,\tau_y) = \mathrm{e}^{-\rho\sqrt{\tau_x^2+\tau_y^2}} \tag{15}$$

is used as a correlation model for images [15,30]. The spatial correlation coefficient ρ is based on the image characteristics, and is specific to the light field. The PSD of the images is computed

by taking the Fourier transform of this auto-correlation function (15). The noise signals n_i are assumed to have a flat power spectrum with noise variance σ_N^2.

Except for the prediction structure, and possibly the geometry error, the factors that determine compression efficiency are fixed for a light field. With a theoretical model, however, it is possible to determine the effects of each of these factors, which may be difficult or impossible to do experimentally. The importance of prediction structure and geometry information can also be studied.

The rate difference performance measure is used to study the effects of these various parameters. It represents the bits per object pixel (bpop) that are saved by using a particular prediction scheme, over simple independent encoding of the images. A more negative rate difference value means better compression efficiency.

The experiments in this section use the real-world *Bust* light field, which consists of 339 images, each of resolution 480×768, and the synthetic *Buddha* light field, which consists of 281 images, each of resolution 512×512. Results for other datasets can be found in [33]. For the theoretical results, the value for spatial correlation $\rho = 0.93$ is estimated from the image data, for both the *Bust* and the *Buddha* light fields. Likewise, the actual camera positions of the light field is used to determine texture shifts.

Theoretical simulation results, showing rate difference performance, are given in Fig. 5, for the *Bust* light field, and Fig. 7, for the *Buddha* light field. Three different encodings are studied. These are illustrated in Figs. 3 and 4, for the *Bust* and *Buddha* light fields, respectively. The first encoding, using independent coding of each image and no prediction between images, serves as the reference. The rate difference of this reference scheme is $\Delta R = 0$. The prediction-based encoding schemes groups images into either two or four levels, where images in the lower levels are used to predict images in levels

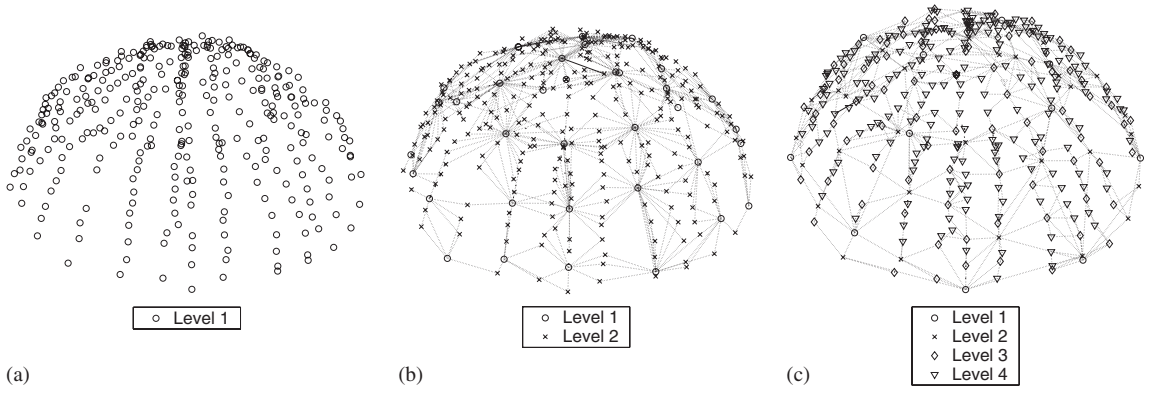

Fig. 3. Prediction structures for the *Bust* light field. (a) No prediction—one level. (b) Prediction—two levels. (c) Prediction—four levels.

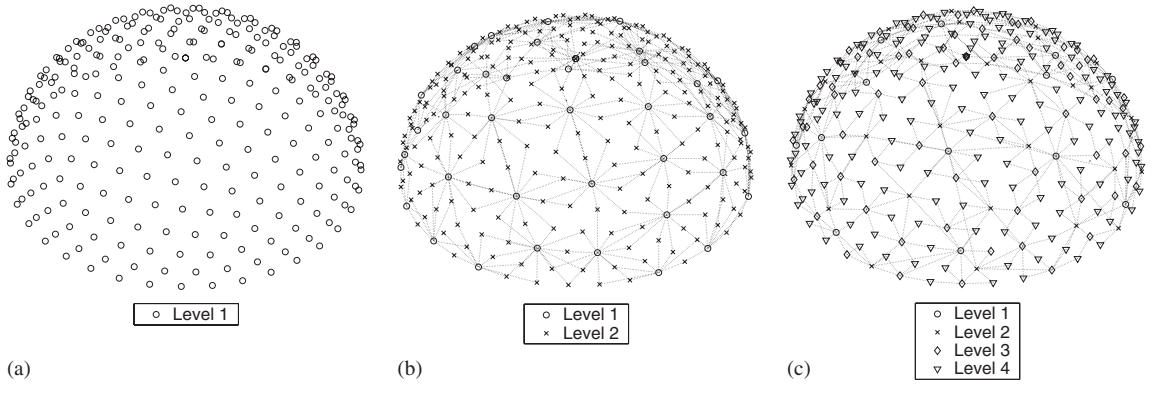

Fig. 4. Prediction structures for the *Buddha* light field. (a) No prediction—one level. (b) Prediction—two levels. (c) Prediction—four levels.

above. The rate difference of the two prediction-based encoding schemes is shown for two levels of geometry accuracy σ_G^2 and a range of independent noise variance values σ_N^2.

In Figs. 5 and 7, there are two levels of geometry error that are shown. The first value of geometry error variance σ_G^2 represents the maximum geometry error variance for reasonable rendering. This value, which assumes a maximum average tolerable shift of 0.5 pixels in the image domain for rendering, depends on the camera directions, and therefore varies from light field to light field. The estimated values are $\sigma_G^2 = 2.0$ for the *Bust* light field and $\sigma_G^2 = 0.5$ for the *Buddha* light field. The other value of geometry error variance $\sigma_G^2 \approx 0$ represents an accurate geometry model. These two levels of geometry error represent the two extremes on the range of geometry error levels to be considered, and the actual geometry error level will likely fall somewhere within this range.

The range of noise variance values σ_N^2 is chosen as to result in a reasonable range of rate differences. Approximately 1 bpop represents the maximum rate savings seen in actual light field coding experiments, described later.

It is clear from Figs. 5 and 7 that accurate geometry, inter-view correlation and prediction can all significantly impact compression results. These effects are inter-dependent. Prediction between images can lead to significant improvements in compression performance, compared to not using prediction. Increasing the number of levels of images in the prediction structure, however, gives only modest gains.

When using prediction, geometry accuracy can have a significant effect on compression efficiency. The rate difference of the curves that use exact geometry is significantly better than that of curves with the less accurate geometry. Also, the graph indicates that a light field that has more inter-view correlation, corresponding to smaller values of σ_N^2, can be encoded much more efficiently.

The combination of these factors is also important. Prediction between images gives high compression efficiency for high inter-view correlation and accurate geometry. For poor geometry and low inter-view correlation, prediction only results in a saving of 0.1–0.2 bpp. The numbers depend on the exact value of inter-view correlation. The bit-rate savings for better geometry information is greater for light fields with high inter-view correlation than for those with low inter-view correlation. More accurate geometry information serves to better exploit the higher correlation between images.

The numerical results obtained from the theory show how the various parameters of a light field coding system affect compression performance. In particular, when there is correlation between images and good geometry accuracy, prediction can significantly improve results. With prediction, a residual image is encoded instead of the original image, resulting in fewer bits used. When more images are predicted, as is the case with more levels of prediction, then the overall compression efficiency improves. Prediction with two levels of images, however, seems to exploit most of inter-view correlation, and there is a small, but limited benefit from using more levels.

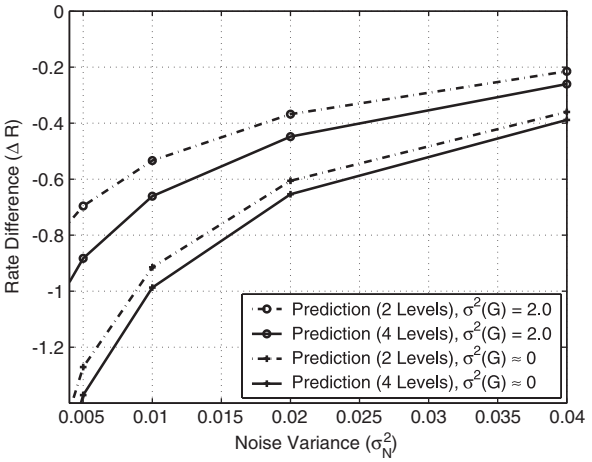

Fig. 5. Theoretical rate difference curves for *Bust* light field.

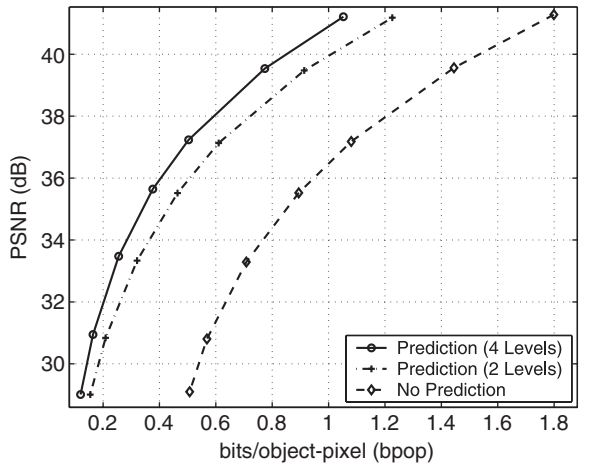

Fig. 6. Experimental rate-PSNR curve for the *Bust* light field.

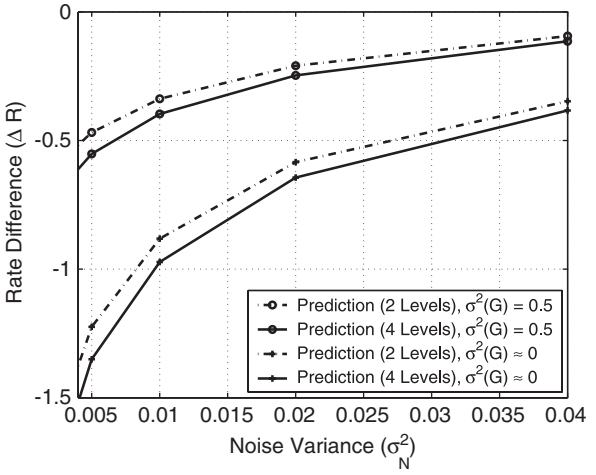

Fig. 7. Theoretical rate difference curves for *Buddha* light field.

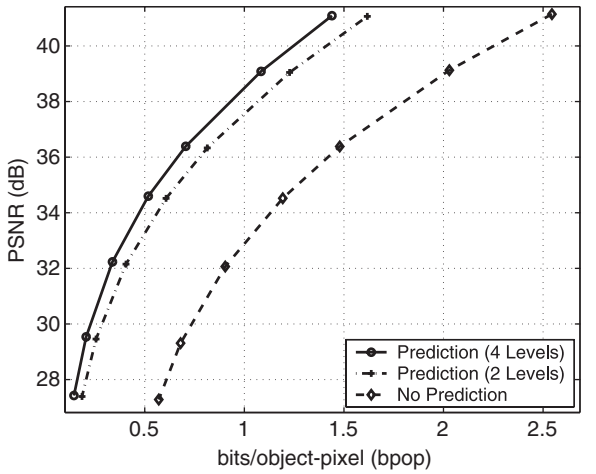

Fig. 8. Experimental rate-PSNR curve for the *Buddha* light field.

Figs. 6 and 8 show the experimental compression results for the same light field datasets using the light field coder in [6]. The light field coder is run at various quality levels, determined by the quantization parameter Q. Values of $3, 4, 6, 8, 12, 20, 31$ are used for Q. The encoding that corresponds to each Q value results in a particular bit-rate, measured by the total number of bits divided by the number of object pixels in the light field (bits per object pixel or bpop), and a distortion value, measured in MSE averaged over all the images in the light field. This average MSE is converted to PSNR and reported in dB.

The three different prediction schemes that are compared in these figures are the same as in the theoretical simulations reported earlier. As expected, the scheme that does not use any prediction gives the poorest compression results, since the correlation between images is not exploited. At high rates, prediction with two levels reduces the bit-rate over no prediction by 0.5 bpop for the *Bust* light field and 0.9 bpop for the *Buddha* light field. This corresponds to an improvement in image quality of more than 3 dB. This corresponds to the improvement predicted by the theoretical simulations. Note that since the *Buddha* dataset is a synthetic and accurate geometry is known, the curves for $\sigma_G^2 \approx 0$ can be used, whereas for the *Bust*, somewhere between the inaccurate and accurate geometry curves. The reduction in bit-rate between using two levels of images, and four levels of images, is much smaller, only about 0.15 bpop for both datasets. This corresponds with the observation from the theoretical results that additional levels of prediction give small but limited improvement in compression performance.

3. Extension to light field streaming

3.1. View-trajectory-dependent rate-distortion

So far, the rate-distortion performance considers only encoding all the images in a light field. This is appropriate for scenarios such as storage or download of an entire light field, where the entire light field can be compressed and decompressed, but is less applicable to scenarios such as streaming where only the necessary or available subset of images is used for rendering. When interactively viewing a light field, views are rendered at a predetermined time, which means that the necessary images must arrive by a particular deadline. Streaming is abstracted as a process which transmits a set of images and, for each rendering instance, makes a particular set of images available. The details about the packet scheduling algorithm and the underlying network conditions are not required for this analysis.

For the streaming scenario, it is appropriate to consider the view-dependent rate-distortion function. The rate is counted over all images that are needed to render the view trajectory, including those that are needed for decoding due to prediction dependencies, in addition to those for rendering. The distortion is measured between the rendered trajectory using the original uncompressed light field, and the rendered trajectory using the reconstructed light field. Using this view-trajectory-dependent rate-distortion measure accounts for

both the compression efficiency of a particular prediction scheme and the random access cost of the prediction dependency.

In order to understand this theoretically, the process of rendering an image must be modeled. A number of reference images may be used to render a particular view. To render a pixel in the novel view, pixels from the reference images are geometry-compensated and linearly combined. Accordingly, a rendered view v_W, for view W, is modeled as a linear combination of the light field images $\{c_i\}$, given by the equation

$$v_W = \sum_{i=1}^{N} K_{W,i} c_i, \tag{16}$$

where $K_{W,i}$ is the rendering weight for view W and image i. The rendering weights are determined by the rendering algorithm, according to factors such as difference in viewing angle [3]. Images that are not used in rendering that view are given the weight of 0. This model only approximates the actual rendering process since a single rendering weight is used per image, instead of per pixel.

Rendering from the reconstructed light field images $\{\hat{c}_i\}$, not all the images in the light field may be available. This may happen often in a streaming scenario where rendering uses whatever images are available at rendering time. In this case, the rendering weights for the view $\{\hat{K}_{W,i}\}$ may be different than those for the rendering from the original images (16). The distorted rendered view $\hat{v}_S$ can be written as

$$\hat{v}_W = \sum_{i=1}^{N} \hat{K}_{W,i} \hat{c}_i. \tag{17}$$

With this rendering model, distortion in the rendered view W due to coding can be calculated as

$$D_W(\phi) = E\{(v_W - \hat{v}_W)^2\} \tag{18}$$

$$= E\left\{ \left(\sum_{i=1}^{N} K_{W,i} c_i - \sum_{i=1}^{N} \hat{K}_{W,i} \hat{c}_i \right)^2 \right\} \tag{19}$$

$$= E\left\{ \left(\sum_{i=1}^{N} K_{W,i} c_i - \sum_{i=1}^{N} \hat{K}_{W,i} c_i + \sum_{i=1}^{N} \hat{K}_{W,i} c_i - \sum_{i=1}^{N} \hat{K}_{W,i} \hat{c}_i \right)^2 \right\} \tag{20}$$

$$= E\left\{ \left(\sum_{i=1}^{N} (K_{W,i} - \hat{K}_{W,i}) c_i + \sum_{i=1}^{N} \hat{K}_{W,i} (c_i - \hat{c}_i) \right)^2 \right\} \tag{21}$$

$$= E\left\{ \left(\sum_{i=1}^{N} (K_{W,i} - \hat{K}_{W,i}) c_i + \sum_{i=1}^{N} \hat{K}_{W,i} (e_i - \hat{e}_i) \right)^2 \right\} \tag{22}$$

$$\approx E\left\{ \left(\sum_{i=1}^{N} (K_{W,i} - \hat{K}_{W,i}) c_i \right)^2 \right\} + E\left\{ \left(\sum_{i=1}^{N} \hat{K}_{W,i} (e_i - \hat{e}_i) \right)^2 \right\} \tag{23}$$

$$\approx E\left\{ \left(\sum_{i=1}^{N} (K_{W,i} - \hat{K}_{W,i}) c_i \right)^2 \right\} + \sum_{i=1}^{N} \hat{K}_{W,i}^2 D_{e_i}(\phi) \tag{24}$$

$$= \sum_{i=1}^{N} \sum_{i=1}^{N} (K_{W,i} - \hat{K}_{W,i})(K_{W,j} - \hat{K}_{W,j}) E\{c_i c_j\} + \sum_{i=1}^{N} \hat{K}_{W,i}^2 D_{e_i}(\phi) \tag{25}$$

$$= \frac{1}{4\pi^2} \int_{\omega_x=-\pi}^{\pi} \int_{\omega_y=-\pi}^{\pi} (\mathbf{K}_W - \hat{\mathbf{K}}_W)^{\mathrm{T}} \cdot \mathbf{\Phi}_{cc} (\mathbf{K}_W - \hat{\mathbf{K}}_W) \, \mathrm{d}\omega_x \, \mathrm{d}\omega_y + \sum_{i=1}^{N} \hat{K}_{W,i}^2 D_{e_i}(\phi). \tag{26}$$

To derive (19), (16) and (17) are substituted into (18). By adding and subtracting the quantity $\sum_{i=1}^{N} \hat{K}_{W,i}^2 c_i$, (20) is obtained, and (21) follows by grouping terms. Assuming closed-loop predictive coding and substituting in (11), (22) is obtained. By assuming the quantization error $e_i - \hat{e}_i$ is uncorrelated with the original signals $c_1, c_2, \ldots, c_n$, the cross-terms are dropped and the result is the

approximation in (23). This assumption is reasonable for smooth probability density functions and uniform quantization, in the high rate regime [1,20]. Under these conditions, and if none of the signals e_i are identical, then it is reasonable to assume that the quantization errors are uncorrelated with each other. Thus, the remaining cross-terms are dropped and the result is the approximation (24). $D_{e_i}(\phi)$ represents the distortion due to coding for residual error i, as defined in (9). The first term in (24) can be expanded to give the expression in (25), which can then be re-written in terms of the known PSD $\mathbf{\Phi}_{cc}$ in (26). $\mathbf{K}_W = [K_{W,1}\ K_{W,2}\ \cdots\ K_{W,N}]^{\mathrm{T}}$ and $\hat{\mathbf{K}}_W = [\hat{K}_{W,1}\ \hat{K}_{W,2}\ \cdots\ \hat{K}_{W,N}]^{\mathrm{T}}$ are the weight vectors.

The two terms in (26) correspond to the two sources of distortion in a rendered view. The first term, consisting of the integral, represents an error-concealment or a "substitution" distortion that results from using a different set of rendering weights in the original and distorted rendered views. If the rendering weights are identical, i.e. $\mathbf{K}_W = \hat{\mathbf{K}}_W$, then this term disappears. The second term represents the contribution of the coding distortion to the rendered distortion. With this simple rendering model, the rendered distortion is simply a linear combination of the coding distortions. Expression (26) can be efficiently calculated numerically, given the rendering weights and the PSD matrices $\mathbf{\Phi}_{cc}$ and $\mathbf{\Phi}_{ee}$.

The rate is calculated over all images that are transmitted to render a particular view or view trajectory. The set of images to render view trajectory $\mathbf{W}$ is denoted as $\tilde{\mathscr{C}}_{\mathbf{W}}$. It includes images that are needed for decoding other images, not just those directly needed for rendering. The rate is

$$R_{\mathbf{W}}(\phi) = \sum_{i \in \tilde{\mathscr{C}}_{\mathbf{W}}} R_{e_i}(\phi). \tag{27}$$

The rate for image i, $R_{e_i}(\phi)$ is defined in (8) and is calculated from the PSD matrix $\mathbf{\Phi}_{ee}$. The distortion for the trajectory,

$$D_{\mathbf{W}}(\phi) = \frac{1}{|\mathbf{W}|} \sum_{t=1}^{|\mathbf{W}|} D_{W_t}(\phi), \tag{28}$$

is simply the average over the distortion for each view W_t, and $|\mathbf{W}|$ is the number of views in trajectory $\mathbf{W}$.

3.2. Streaming simulation and experimental results

The rate-distortion performance for a streaming session can be computed from the theoretical model. The rate-distortion performance for streaming depends upon which images are transmitted, and which images are used for rendering. The interactive light field streaming system of [37,35] selects images for transmission so as to optimize rendered image quality, with constraints and knowledge about the network. For low to medium rate scenarios, only a small subset of the images required for rendering may be transmitted. Streaming traces can provide the images that were transmitted, whether they were received, and, if so, when they were received, all for a given streaming session. This can determine the images used for rendering. The theoretical rate can be calculated based on the images that the trace indicates are transmitted. The rendered distortion can be calculated by first determining the rendering weight vectors for the available images, and then calculating the theoretical distortion.

Several parameters need to be set for each light field used in the theoretical derivation, as described in Section 3.1. These include the noise variance σ_{N}^2, geometry error variance σ_{G}^2, and the spatial correlation ρ. The values that are used for the *Bust* dataset are $\sigma_{\mathrm{N}}^2 = 0.015$, $\sigma_{\mathrm{G}}^2 \approx 0$, and $\rho = 0.93$, and for the *Buddha* dataset are $\sigma_{\mathrm{N}}^2 = 0.01$, $\sigma_{\mathrm{G}}^2 \approx 0$, and $\rho = 0.93$. The values chosen for σ_{G}^2 and σ_{N}^2 were guided by the discussion in Section 2.4. The rate-distortion trade-off parameter ϕ_i must correspond to the quantization parameter. In the experiments, a quantization step size of $Q = 3$ was used. The image pixel values range from 0 to 255. Assuming an image variance of 1000 leads to a quantization distortion of $D_i(\phi_i) \approx \phi_i \approx 0.003$.

Figs. 9 and 10 show the simulation results comparing the streaming performance for three different encodings of the light field, for two different types of user viewing trajectories, for the *Bust* and *Buddha* datasets. These figures also show the experimental results side-by-side for comparison. The three different encodings of the light field are: INTRA, or independent coding; prediction with two levels of images and; prediction with four levels of images. The two different types of trajectories are denoted *slow* and *fast*. Each trajectory consists of 50 views, rendered every 50 ms, in total constituting a total time of 2.5 s. The *slow* trajectories represent deliberate, predictable movement by the user, and cover only a small portion of the viewing hemisphere, while *fast* trajectories represent erratic viewing that tends to access more of the viewing hemisphere.

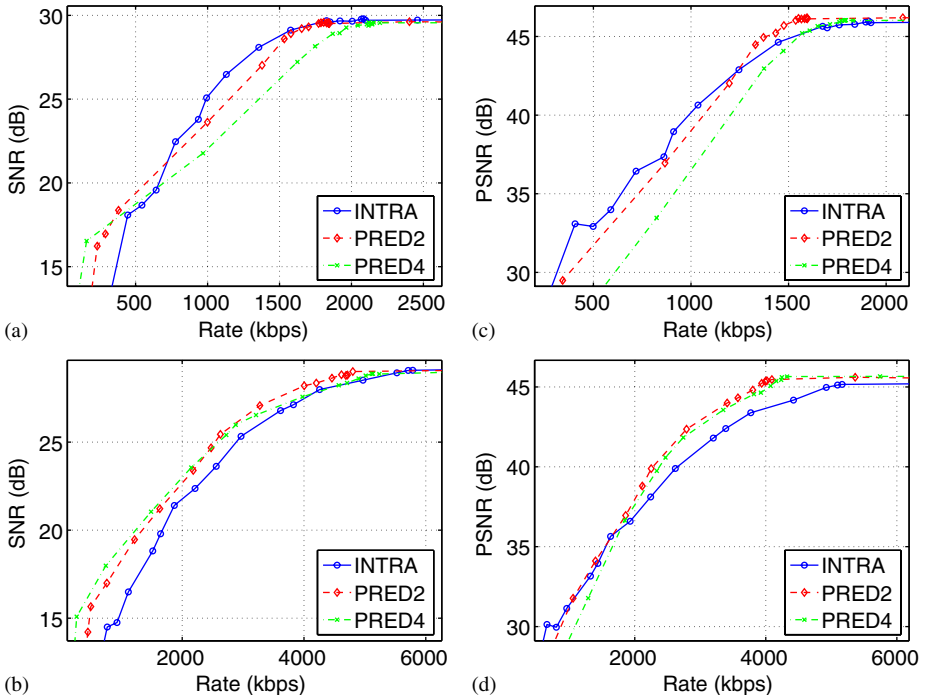

Fig. 9. Theoretical and experimental streaming results for the *Bust* light field, comparing three different encodings. Receiver-driven rate-distortion optimized streaming is used. (a) Theoretical—*slow* trajectory, (b) theoretical—*fast* trajectory, (c) experimental—*slow* trajectory, (d) experimental—*fast* trajectory.

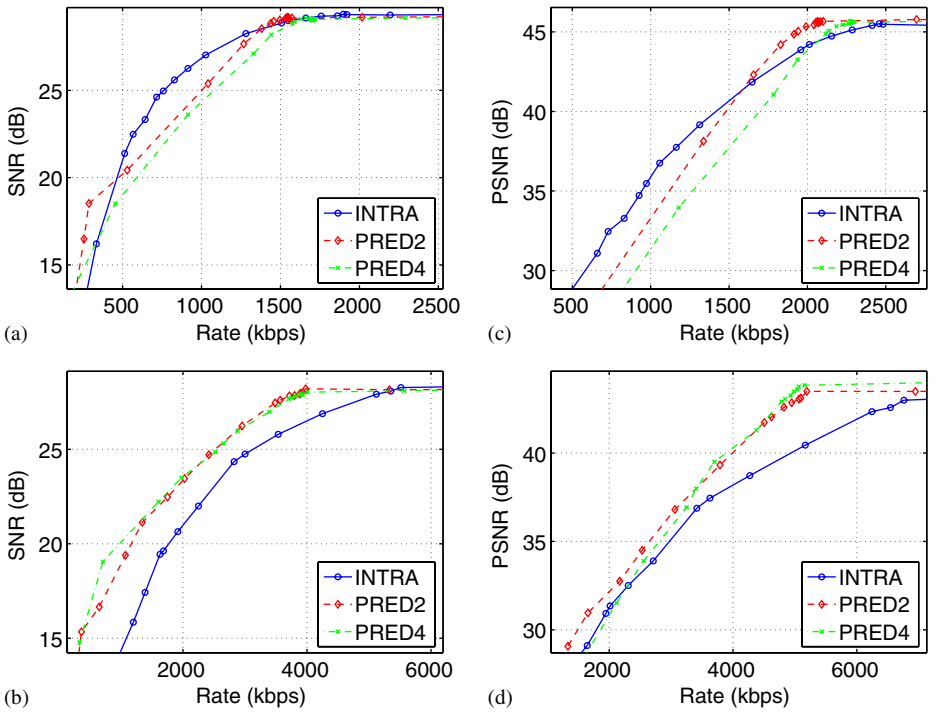

Fig. 10. Theoretical and experimental streaming results for the *Buddha* light field, comparing three different encodings. Receiver-driven rate-distortion optimized streaming is used. (a) Theoretical—*slow* trajectory, (b) theoretical—*fast* trajectory, (c) experimental—*slow* trajectory, (d) experimental—*fast* trajectory.

Results are averaged over 10 trials of 10 trajectories of each type. The experimental streaming results, that deal with pixel values from 0 to 255, are reported as PSNR in dB, while the theoretical results, that deal with a Gaussian random process, use SNR, also in dB. These two measures can be related with a vertical shift. The distortion values are calculated by averaging the SNR or PSNR distortion values for all the views in the trajectory.

The simulations show the same trends as the experimental results. As the trajectory changes from the *slow* to *fast*, the performance of INTRA encoding degrades, and the performance of the prediction-based encodings improves. The *fast* trajectory covers more of the viewing hemisphere than the *slow* trajectory and, typically, requires more images. Prediction-based encodings, with a fixed prediction structure, can do better when the required images match those provided by prediction structure. The INTRA encoding is superior to the other trajectories only for the *slow* trajectories, and some range of bit-rates, in both the experimental and theoretical results. More details about the light field streaming system can be found in [37,38]. A comprehensive set of simulation results, including those for other datasets, can be found in [33].

While the simulation results have similar to that of the actual experimental results, they are clearly not identical. This can be attributed to the numerous simplifying assumptions about the 3-D geometry, image formation, and the statistics of the images. Despite these assumptions, these experiments indicate that the theoretical framework can reasonably explain actual streaming results. This theoretical framework could be used, for instance, to better design coding prediction dependencies or scheduling algorithms. In [36,33], the theoretical rate-distortion framework is found to be a useful tool in estimating distortion values when doing rate-distortion optimized packet scheduling. In that work, the theoretical framework is used in a hybrid manner together with samples of actual distortion values. The resulting estimated distortion allows the system to achieve nearly the same rate-distortion streaming performance as when using actual distortion values [36,33].

4. Conclusions

A theoretical framework to analyze the rate-distortion performance of a light field coding and streaming system is proposed. In the framework, the encoding and streaming performance is affected by various parameters, such as the correlation within an image and between images, geometry accuracy and the prediction dependency structure. Using the framework, the effects of these parameters can be isolated and studied. Specifically, the framework shows that compression performance is significantly affected by three main factors, namely, geometry accuracy, correlation between images, and the prediction dependency structure. Moreover, the effects of these factors are inter-dependent. Prediction is only useful when there is both accurate geometry and sufficient correlation between images. The gains due to geometry accuracy and image correlation are not simply additive. The larger the correlation between images, the greater the benefit of more accurate geometry over less accurate geometry. In converse, with low correlation between images, there is little benefit to any level of accuracy in the geometry. In the datasets studied, the theory indicates that most of the benefit of prediction is realized with only two levels of images in the prediction dependency structure. This result is confirmed by actual experimental results.

The theoretical framework for light field coding is extended and used to study the performance of streaming compressed light fields. In order to extend the framework, a view-trajectory-dependent rate-distortion function is derived.

The derivation shows that the distortion is composed to two additive parts: distortion due to coding, and distortion resulting from using only a subset of the required images for rendering. Theoretical simulation results, using actual streaming traces, reveal that the prediction structure that gives the best streaming performance depends heavily upon the desired viewing trajectory. Independent coding of the light field images is, in fact, the best trajectory for certain view trajectories of certain datasets. This observation is confirmed with actual streaming experimental results. The simulation results, in general, show similar trends and characteristics to the actual experimental streaming results. There is not an exact match, due the numerous simplifications and assumptions made in the model to make the analysis tractable. The framework, despite these limitations, has been effectively used as part of the procedure for estimating distortion values for rate-distortion optimized packet scheduling.

Acknowledgment

The authors would like to thank Markus Flierl for his valuable comments on this manuscript.

References

[1] W.R. Bennett, Spectra of quantized signals, Bell System Technical J. 27 (July 1948).
[2] T. Berger, Rate Distortion Theory: A Mathematical Basis for Data Compression, Prentice-Hall, Englewood Cliffs, NJ, 1971.
[3] C. Buehler, M. Bosse, L. McMillan, S. Gortler, M. Cohen, Unstructured lumigraph rendering, in: Computer Graphics (Proceedings of SIGGRAPH01), 2001, pp. 425–432.
[4] J.-X. Chai, X. Tong, S.-C. Chan, H.-Y. Shum, Plenoptic sampling, in: Computer Graphics (Proceedings of SIGGRAPH00), August 2000, pp. 307–318.
[5] C.-L. Chang, X. Zhu, P. Ramanathan, B. Girod, Inter-view wavelet compression of light fields with disparity-compensated lifting, in: Proceedings of the SPIE Visual Communications and Image Processing VCIP-2003, Lugano, Switzerland, July 2003.
[6] C.-L. Chang, X. Zhu, P. Ramanathan, B. Girod, Shape adaptation for light field compression, in: Proceedings of the IEEE International Conference on Image Processing ICIP-2003, Barcelona, Spain, September 2003.
[7] N. Farvardin, J.W. Modestino, Rate-distortion performance of DPCM schemes for autoregressive sources, IEEE Trans. Image Process. 31 (3) (May 1985) 402–418.
[8] M. Flierl, B. Girod, Multihypothesis motion estimation for video coding, in: Proceedings of the Data Compression Conference 2001, Snowbird, UT, USA, March 2001.
[9] M. Flierl, B. Girod, Multihypothesis motion-compensated prediction with forward adaptive hypothesis switching, in: Proceedings of the Picture Coding Symposium 2001, Seoul, Korea, April 2001.
[10] M. Flierl, B. Girod, Video coding with motion compensation for groups of pictures, in: Proceedings of the IEEE International Conference on Image Processing ICIP-2002, vol. I, Rochester, NY, USA, September 2002, pp. 69–72.
[11] M. Flierl, B. Girod, Investigation of motion-compensated lifted wavelet transforms, in: Proceedings of the Picture Coding Symposium 2003, Saint-Malo, France, April 2003.
[12] M. Flierl, B. Girod, Video coding with motion-compensated lifted wavelet transforms, EURASIP J. Image Comm. Special Issue on Subband/Wavelet Interframe Video Coding 19 (7) (August 2004) 561–575.
[13] R. Gallager, Information Theory and Reliable Communication, Wiley, New York, 1968.
[14] A. Gersho, R.M. Gray, Vector Quantization and Signal Compression, Kluwer Academic Publishers, Dordrecht, 1992.
[15] B. Girod, The efficiency of motion-compensating prediction for hybrid coding of video sequences, IEEE J. Selected Areas Comm. SAC-5 (August 1987) 1140–1154.
[16] B. Girod, Motion-compensating prediction with fractional pel accuracy, IEEE Trans. Comm. 41 (April 1993) 604–612.
[17] B. Girod, Efficiency analysis of multihypothesis motion-compensated prediction for video coding, IEEE Trans. Image Process. 9 (2) (February 2000) 173–183.
[18] B. Girod, C.-L. Chang, P. Ramanathan, X. Zhu, Light field compression using disparity-compensated lifting, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 2003, vol. IV, Hong Kong, China, April 2003, pp. 761–764.
[19] S.J. Gortler, R. Grzeszczuk, R. Szeliski, M.F. Cohen, The lumigraph, in: Computer Graphics (Proceedings of SIGGRAPH96), August 1996, pp. 43–54.
[20] R.M. Gray, D.L. Neuhoff, Quantization, IEEE Trans. Inform. Theory 44 (6) (October 1998) 2325–2383.
[21] A. Isaksen, L. McMillan, S.J. Gortler, Dynamically reparameterized light fields, in: Computer Graphics (Proceedings of SIGGRAPH00), August 2000.
[22] N.S. Jayant, P. Noll, Digital Coding of Waveforms, Prentice-Hall, Englewood Cliffs, NJ, 1984.
[23] A.N. Kolmogorov, On the Shannon theory of information transmission in the case of continuous signals, IEEE Trans. Image Process. 2 (4) (December 1956) 102–108.
[24] M. Levoy, P. Hanrahan, Light field rendering, in: Computer Graphics (Proceedings of SIGGRAPH96), August 1996, pp. 31–42.
[25] M. Levoy, K. Pulli, et al., The digital Michelangelo project: 3D scanning of large statues, in: Computer Graphics (Proceedings of SIGGRAPH00), August 2000, pp. 131–144.
[26] M. Magnor, P. Eisert, B. Girod, Model-aided coding of multi-viewpoint image data, in: Proceedings of the IEEE International Conference on Image Processing ICIP-2000, vol. 2, Vancouver, Canada, September 2000, pp. 919–922.
[27] M. Magnor, A. Endmann, B. Girod, Progressive compression and rendering of light fields, in: Proceedings of the Vision, Modelling and Visualization 2000, November 2000, pp. 199–203.
[28] M. Magnor, B. Girod, Data compression for light field rendering, IEEE Trans. Circuits Systems Video Technol. 10 (3) (April 2000) 338–343.
[29] M. Magnor, P. Ramanathan, B. Girod, Multi-view coding for image-based rendering using 3-D scene geometry, IEEE Trans. Circuits Systems Video Technol. 13 (11) (November 2003) 1092–1106.
[30] J.B. O'Neal Jr., T. Raj Natarajan, Coding isotropic images, IEEE Trans. Inform. Theory 23 (6) (November 1977) 697–707.
[31] I. Peter, W. Strasser, The wavelet stream: progressive transmission of compressed light field data, in: IEEE Visualization 1999 Late Breaking Hot Topics, October 1999, pp. 69–72.
[32] M.S. Pinsker, in: Trudy, Third All-Union Mathematical Conference, vol. 1, 1956, p. 125.
[33] P. Ramanathan, Compression and interactive streaming of light fields, Ph.D. Thesis, Stanford University, Stanford, CA, 2005.
[34] P. Ramanathan, M. Flierl, B. Girod, Multi-hypothesis disparity-compensated light field compression, in: Proceedings of the IEEE International Conference on Image Processing ICIP-2001, October 2001.
[35] P. Ramanathan, B. Girod, Rate-distortion optimized streaming of compressed light fields with multiple representations, in: Packet Video Workshop 2004, Irvine, CA, USA, December 2004.

[36] P. Ramanathan, B. Girod, Receiver-driven rate-distortion optimized streaming of light fields, in: Proceedings of the IEEE International Conference on Image Processing ICIP-2005, Genoa, Italy, September 2005.

[37] P. Ramanathan, M. Kalman, B. Girod, Rate-distortion optimized streaming of compressed light fields, in: Proceedings of the IEEE International Conference on Image Processing ICIP-2003, vol. 3, Barcelona, Spain, September 2003, pp. 277–280.

[38] P. Ramanathan, E. Steinbach, P. Eisert, B. Girod, Geometry refinement for light field compression, in: Proceedings of the IEEE International Conference on Image Processing ICIP-2002, vol. 2, Rochester, NY, USA, September 2002, pp. 225–228.

[39] X. Tong, R.M. Gray, Interactive rendering from compressed light fields, IEEE Trans. Circuits Systems Video Technol. 13 (11) (November 2003) 1080–1091.

[40] D.N. Wood, D.I. Azuma, K. Aldinger, B. Curless, T. Duchamp, D.H. Salesin, W. Stuetzle, Surface light fields for 3D photography, in: Computer Graphics (Proceedings of SIGGRAPH00), August 2000, pp. 287–296.

[41] C. Zhang, T. Chen, Spectral analysis for sampling image-based rendering data, IEEE Trans. Circuits Systems Video Technol. 13 (11) (November 2003) 1038–1050.

[42] C. Zhang, J. Li, Compression of lumigraph with multiple reference frame (MRF) prediction and just-in-time rendering, in: Proceedings of the Data Compression Conference 2000, Snowbird, UT, USA, March 2000, pp. 253–262.