THEORETICAL ANALYSIS OF THE RATE-DISTORTION PERFORMANCE OF A LIGHT FIELD STREAMING SYSTEM

Prashant Ramanathan and Bernd Girod
Information Systems Laboratory
Department of Electrical Engineering
Stanford University
{pramanat,bgirod}@Stanford.EDU

***Abstract*— Large image-based rendering data sets such as light fields require efficient compression and random access to individual images for applications such as interactive streaming to a remote user. In our earlier work, we propose a theoretical framework to analyze the trade-off between compression efficiency and random access. In this current paper, we extend the theoretical framework by calculating distortion for a more general rendering scenario. We then use this result to derive the theoretical rate-distortion function for a streaming session, and show that the simulation results have similar trends as the actual experimental results.**

I. INTRODUCTION

A *light field* [1], [2] is an image-based rendering data set that can be used for interactive photorealistic rendering of 3-D objects and scenes. A light field is a 4-D data set which is often parameterized as a 2-D array of images. In this paper, we use a 2-D hemispherical arrangement of cameras surrounding the object of interest in the light field. This data set represents the outgoing radiance from a particular scene or object at all points in 3-D space and in all directions. Novel views of the object or scene can be created by appropriately sampling the pixels in the captured images of the light field.

The scenario considered in this paper is one where the light field data set is accessed by a remote user over a network such as the Internet. Since light field data sets are very large, efficient compression of the data is necessary. Even with state-of-the-art compression, however, downloading the entire data set is inconvenient. It is therefore more attractive to stream the image data to the user while they interact with it.

There has been significant work on the compression of light fields and related data sets, such as concentric mosaics [4]. Approaches include vector quantization [1], [4], [5], MPEG-style DCT-transform-based encoding using prediction between images [6]–[9], and wavelet-based approaches [6], [10], including those using disparity-compensated lifting [11], [12]. In this work, we focus on an algorithm that predicts between images. The basic idea of this algorithm is described in [6] and more implementation details are given in [13], [14].

In [15], a rate-distortion optimized streaming framework is proposed for compressed light fields. The framework takes into account network conditions, history of transmissions and acknowledgments, prediction dependency structure and the size and importance of each data unit. Using this information, the algorithm selects which images to send at each transmission opportunity so as to minimize the distortion that the remote user experiences, given a rate constraint over the network.

Many parameters affect the rate-distortion streaming performance including the prediction dependency structure, which determines coding efficiency and random access to images; the accuracy of the geometry information which is used in coding, rendering and decoding; and the streaming scheduler, which affects which images that are available for rendering at the remote user. In this paper, we propose a theoretical framework that accounts for these different parameters, and which can be used to analyze the rate-distortion performance of coding, rendering and streaming system.

In Section II, we summarize the prior work upon which this paper is based. In Section III, we extend the framework so that it can be used to analyze a streaming scenario. Results from theoretical simulations and a comparison with experimental results are described in Section IV.

II. PRIOR WORK

We start by summarizing the signal model presented in [3], [6], [16]. Figure 1 shows a block diagram of this model of the light field coder. Geometry-compensated light field images $\{c_i\}$ are produced from a hypothetical texture image v through a random shift (modelling inaccuracy in our knowledge of the geometry) and addition of independent noise (modelling non-Lambertian effects and camera noise).

The coding of the light field is modelled by a linear transform T, which attempts to decorrelate the signals,

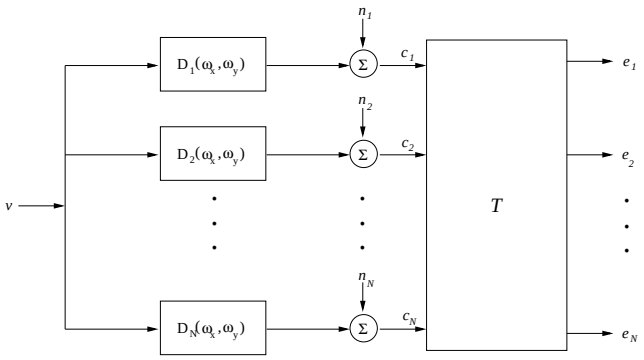


Fig. 1. The signals c_i are produced by shifting the original texture signal v by (Δ_{xi}, Δ_{yi}) (described by the transfer function $D_i(\omega_x, \omega_y)$) and adding signal-independent noise n_i. The signals $\{c_i\}$ are then encoded by first linearly transforming them with matrix T to give the signals $\{e_i\}$. These signals are then independently encoded.

followed by independent encoding of the tranformed signals. Different transforms T represent different methods for predicting an image from other images. By considering the texture and noise images to be jointly wide-sense stationary, we can derive a power spectral density (PSD) function for the transformed images.

If we additionally consider these images to be jointly Gaussian, we can derive their rate-distortion function. For each image e_i, we can obtain the rate $R(\phi)$ and mean-squared error distortion $D(\phi)$ in terms of the rate-distortion trade-off parameter ϕ, using the power spectral density $\Phi_{e_i e_i}$ of signal e_i [17].

In [3], we use these quantities to derive the rate-distortion function for a given viewing trajectory in the following manner. We model a rendered view v_S as a linear combination of the light field images $\{c_i\}$, given by the equation $v_S = \sum_{i=1}^{N} K_{S,i} c_i$ where $\{K_{S,i}\}$ are the rendering weights. Using the reconstructed light field images $\{\hat{c}_i\}$, we obtain the distorted rendered view $\hat{v}_S = \sum_{i=1}^{N} K_{S,i} \hat{c}_i$. The constants $K_{S,i}$ depend on the rendering algorithm.

Assuming high rate we derive in [3] the distortion in the rendered novel view due to quantization as

$$D_S(\phi) = \sum_{i=1}^{N} K_{S,i}^2 D_i(\phi). \tag{1}$$

The rate is calculated by summing over all images needed for rendering or prediction, as determined by the prediction dependency structure. In the next section, we show how to extend this analysis to a streaming scenario.

III. RATE-DISTORTION ANALYSIS FOR STREAMING

A typical streaming system consists of a server that contains pre-compressed image data sending the appropriate images to a remote client that renders the desired view from these images. The network that connects the server and remote client has limited bandwidth and may sometimes lose images, or delay their transmission.

For interactive rendering, where views must be rendered according to strict deadlines to preserve interactivity, it may be necessary to render without the ideal set of images. We extend our analysis above to this situation. If certain images are not available for rendering, then the rendering algorithm substitutes the closest available images. We render only with images that are correctly decoded. This gives us the rendered image $\hat{v_S} = \sum_{i=1}^{N} \tilde{K}_{S,i} \hat{c}_i$, which uses the new rendering weights $\tilde{K}$ instead of the ideal weights K.

As before, we compare with the rendering from the uncompressed data set $v_S = \sum_{i=1}^{N} K_{S,i} c_i$. Assuming high rate, as in our prior work, we derive the approximation for the distortion in (2)-(10).

To derive (3), the definitions of v_S and $\hat{v}_S$ are substituted into (2). By adding and subtracting the quantity $\Sigma_{i=1}^{N} \hat{K}_{S,i} c_i$, (4) is obtained, and (5) follows by re-grouping terms. By assuming closed-loop predictive coding, and consequently recognizing that $c_i - \hat{c}_i = e_i - \hat{e}_i$, (6) is obtained. The next two steps require the high-rate assumption. At high rates, the quantization errors $e_i - \hat{e}_i$ are uncorrelated with the original signals $c_1, c_2, ..., c_N$, which allows for their cross-terms to be dropped, resulting in (7). As well, at high rates, the quantization errors become uncorrelated with each other, which allows yet more cross-terms to be dropped, and results in the approximation (8). $D_i(\phi)$ represents the distortion due to coding for residual error image e_i. The first term in (8) can be expanded to give the expression in (9), which can then be re-written in terms of the known PSD $\mathbf{\Phi}_{cc}$ of $\mathbf{c}$ in (10). $\mathbf{K}_S = [K_{S,1}\ K_{S,2}\ \cdots\ K_{S,N}]^T$ and $\hat{\mathbf{K}}_S = [\hat{K}_{S,1}\ \hat{K}_{S,2}\ \cdots\ \hat{K}_{S,N}]^T$ are the weight vectors.

The second term in (10) can be identified as the distortion due to quantization that we saw in (1). The first term in (10) can be thought of as the "substitution" distortion, resulting from not using the ideal images. Note that if $\mathbf{K}_S = \hat{\mathbf{K}}_S$, then we obtain the expression in (1).

With (10), we can theoretically derive the distortion that the remote user experiences when rendering using a particular set of available images, which may or may not be the ideal set of images for rendering. If the desired view-point and the available and ideal set of images are known for a streaming session, then the theoretical distortion can be computed at each rendering instance. Likewise, if the images that are transmitted for a particular

$$
\begin{aligned}
D_S(\phi) &= E\{(v_S - \hat{v}_S)^2\} && (2)\\
&= E\{(\Sigma_{i=1}^N K_{S,i} c_i - \Sigma_{i=1}^N \hat{K}_{S,i} \hat{c}_i)^2\} && (3)\\
&= E\{(\Sigma_{i=1}^N K_{S,i} c_i - \Sigma_{i=1}^N \hat{K}_{S,i} c_i + \Sigma_{i=1}^N \hat{K}_{S,i} c_i - \Sigma_{i=1}^N \hat{K}_{S,i} \hat{c}_i)^2\} && (4)\\
&= E\{(\Sigma_{i=1}^N (K_{S,i} - \hat{K}_{S,i}) c_i + \Sigma_{i=1}^N \hat{K}_{S,i} (c_i - \hat{c}_i))^2\} && (5)\\
&= E\{(\Sigma_{i=1}^N (K_{S,i} - \hat{K}_{S,i}) c_i + \Sigma_{i=1}^N \hat{K}_{S,i} (e_i - \hat{e}_i))^2\} && (6)\\
&\approx E\{(\Sigma_{i=1}^N (K_{S,i} - \hat{K}_{S,i}) c_i)^2\} + E\{(\Sigma_{i=1}^N \hat{K}_{S,i} (e_i - \hat{e}_i))^2\} && (7)\\
&\approx E\{(\Sigma_{i=1}^N (K_{S,i} - \hat{K}_{S,i}) c_i)^2\} + \Sigma_{i=1}^N \hat{K}_{S,i}^2 D_{e_i}(\phi) && (8)\\
&= \Sigma_{i=1}^N \Sigma_{j=1}^N (K_{S,i} - \hat{K}_{S,i})(K_{S,j} - \hat{K}_{S,j}) E\{c_i c_j\} + \Sigma_{i=1}^N \hat{K}_{S,i}^2 D_{e_i}(\phi) && (9)\\
&= \frac{1}{4\pi^2} \int_{\omega_x=-\pi}^{\pi} \int_{\omega_y=-\pi}^{\pi} (\mathbf{K}_S - \hat{\mathbf{K}}_S)^T \, \mathbf{\Phi}_{cc} \, (\mathbf{K}_S - \hat{\mathbf{K}}_S) \, d\omega_x \, d\omega_y + \Sigma_{i=1}^N \hat{K}_{S,i}^2 D_{e_i}(\phi). && (10)
\end{aligned}
$$

streaming session are known, then the theoretical rate can be computed using rates $R_i(\phi)$.

We can obtain this information easily from traces taken from an actual streaming simulation. This is the approach taken in the next section where the theoretical and experimental rate-distortion curves are compared. The purpose is to show that the theoretical model and related equations to compute rate-distortion performance give results that are similar to actual experimental results.

IV. RESULTS

In our experiments, we simulate rate-distortion optimized streaming over a lossy packet network [15]. The traces from these simulations record the images that were transmitted, as well as the images available for each rendering time instance. This information can be used to derive rate-distortion streaming performance theoretically, which can then be compared to the actual experimental results.

Figures 2 and 3 shows the experimental and theoretical rate-distortion streaming results for the *Bust* and *Horse* light fields, respectively, averaged over 10 random viewing trajectories. Each viewing trajectory consists of 25 views around the camera, rendered once every 100 ms. Two encodings are considered: hierarchical predictive encoding of the images and independent (INTRA) coding [6].

The transmission bit rate is given in terms of kilobits per second (kbps). For the theoretical results, we sum over bit rates for all transmitted images, measured in bits per object pixel, and multiply by a factor that accounts for the ratio of rendered pixels to object pixels and the total transmission time, to convert this to transmission rate. We should note that the high rate assumption only applies to the coding of the residual images. Thus, even with this assumption, we can obtain theoretical results for a wide range of transmission rates.

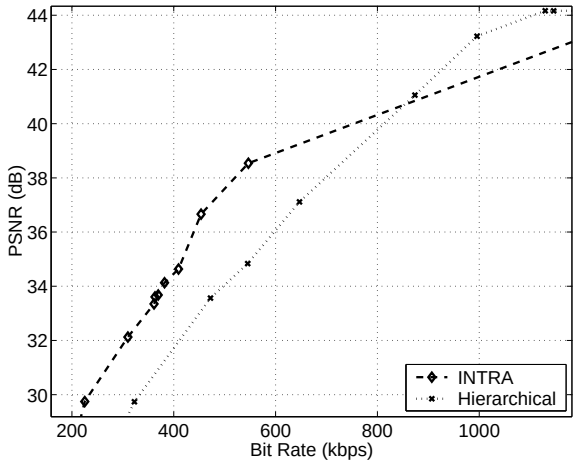
For the experimental results, the distortion is given in terms of PSNR, whereas for the theoretical results, the distortion is stated in terms of SNR. SNR can be related to PSNR by shifting the entire graph vertically, by an amount given by the variance of the images, in our case, a shift of approximately 14 dB.

The experimental results show the relative streaming performance of hierarchical predictive encoding and INTRA coding, encoded at the highest rate, or finest quantization level $Q = 3$. For the *Bust* data set, both schemes perform similarly, but with INTRA slightly outperforming hierarchical coding for much of the range of bit-rates. For the *Horse* data set, INTRA coding is significantly better than hierarchical coding for most of the range of bit-rates.

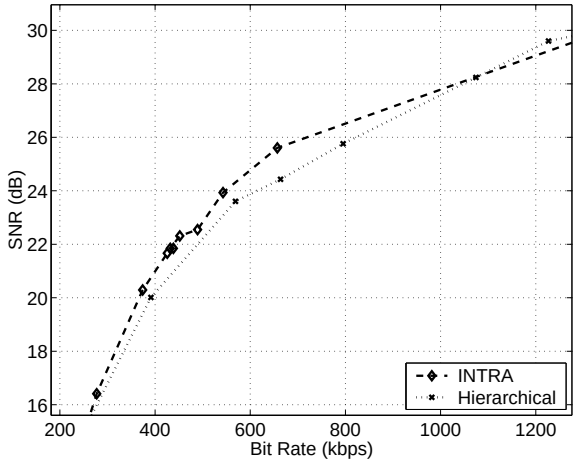
The trends in our theoretical results are similar to those of our experimental results. To calculate $R_i(\phi)$ and $D_i(\phi)$ in our simulations, a value of $\phi = 0.001$ is used. This approximately corresponds to the quantization level $Q = 3$ in our real light field coder. At high transmission rates, all images are sent and thus we are limited in image quality by the quantization distortion. As the rate is decreased, not all needed images are sent, and the substitution distortion dominates.

For the *Bust* data set, we observe a difference of 9 dB in SNR between the transmission rates of 400 and 1200 kbps. Our experimental results shows a corresponding difference of $10 - 12$ dB in PSNR for the same range of bit rates. We also observe similar results for the *Horse* data set. In terms of the comparison between INTRA coding and hierarchical coding, our theoretical results indicate that INTRA coding is slightly better than hierarchical coding for the *Bust* data set, and INTRA coding is significantly better than hierarchical coding for the *Horse* data set.

Because of the numerous simplifications in the theoretical model, however, the theoretical results cannot accurately predict the magnitude of the improvement between the different prediction schemes. Nevertheless, we see that our relatively simple theoretical analysis can capture how the rendered view distortion is affected by rendering with other than the ideal set of images. Presently, this analysis

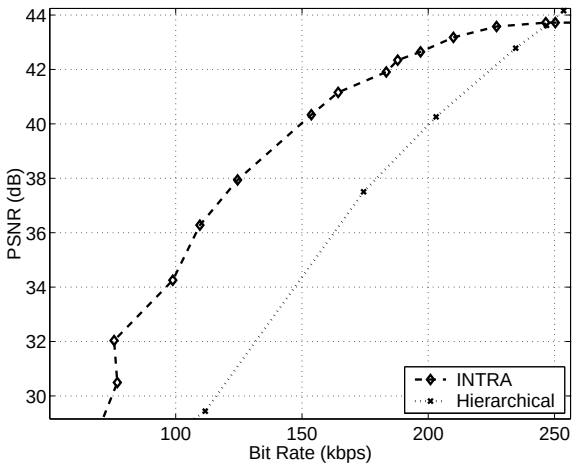


(a) experimental results

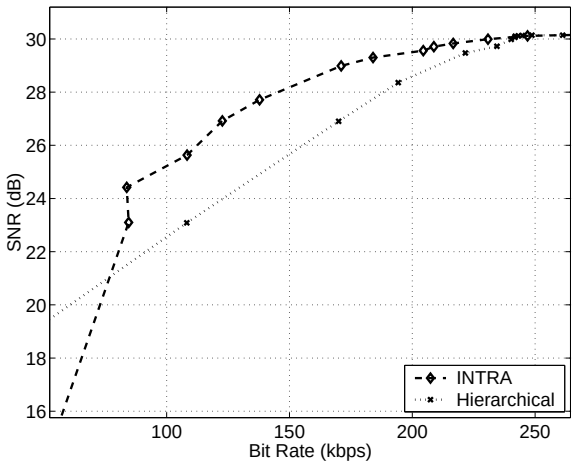


(b) theoretical results

Fig. 2. Experimental and theoretical rate-distortion streaming performance for the *Bust* light field, averaged over 10 random trajectories. Hierarchical coding and INTRA coding have similar streaming performance.



(a) experimental results



(b) theoretical results

Fig. 3. Experimental and theoretical rate-distortion streaming performance for the *Horse* light field, averaged over 10 random trajectories. INTRA coding significantly outperforms hierarchical coding.

still depends on traces of simulated streaming sessions. To allow the study of the various parameters of interest in the light field coding, rendering and streaming system would require modelling and abstracting the streaming system as well. This remains as future work.

V. CONCLUSIONS

Our previous work derived a theoretical view-trajectory-dependent rate-distortion function of a light field. In this paper, we extended this previous work by calculating distortion for the scenario where all the required images may not be available for rendering. We showed that, for this scenario, the distortion is the sum of two terms: the quantization distortion due to coding, and a "substitution" distortion that results from not using all required images for rendering. We use this result to compute the theoretical rate-distortion streaming performance, given traces from simulated streaming sessions. We find that the theoretical results show the same trends as the experimental results. Thus, our theoretical model captures the effects of quantization distortion and "substitution" distortion of the rendered views in a streaming system.

ACKNOWLEDGMENT

This work has been supported by NSF Grant No. ECS-0225315.

REFERENCES

[1] M. Levoy and P. Hanrahan, "Light field rendering," in *Computer Graphics (Proc. SIGGRAPH96)*, August 1996, pp. 31–42.

[2] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Computer Graphics (Proc. SIGGRAPH96)*, August 1996, pp. 43–54.

[3] P. Ramanathan and B. Girod, "Rate-distortion analysis of random access for compressed light fields," in *Proc. IEEE Intl. Conf. on Image Processing ICIP-2004*, 2004, accepted.

[4] H.-Y. Shum and L.-W. He, "Rendering with concentric mosaics," in *Computer Graphics (Proc. SIGGRAPH99)*, August 1999, pp. 299–306.

[5] X. Tong and R. M. Gray, "Interactive rendering from compressed light fields," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, no. 11, pp. 1080–1091, November 2003.

[6] M. Magnor, P. Ramanathan, and B. Girod, "Multi-view coding for image-based rendering using 3-D scene geometry," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, no. 11, pp. 1092–1106, November 2003.

[7] M. Magnor and B. Girod, "Data compression for light field rendering," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 10, no. 3, pp. 338–343, April 2000.

[8] C. Zhang and J. Li, "Compression of lumigraph with multiple reference frame (MRF) prediction and just-in-time rendering," in *Proc. of the Data Compression Conference 2000*, 2000, pp. 253–262.

[9] ——, "Compression and rendering of concentric mosaics with reference block codec (RBC)," in *Proc. SPIE Visual Comm. and Image Processing VCIP-1999*, 2000, pp. 43–54.

[10] L. Luo, Y. Wu, J. Li, and Y.-Q. Zhang, "Compression of concentric mosaic scenery with alignment and 3D wavelet transform," in *Proc. SPIE Visual Comm. and Image Processing VCIP-1999*, 2000, pp. 89–100.

[11] B. Girod, C.-L. Chang, P. Ramanathan, and X. Zhu, "Light field compression using disparity-compensated lifting," in *Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing 2003*, vol. IV, Hong Kong, China, April 2003, pp. 761–764.

[12] C.-L. Chang, X. Zhu, P. Ramanathan, and B. Girod, "Inter-view wavelet compression of light fields with disparity-compensated lifting," in *Proc. SPIE Visual Comm. and Image Processing VCIP-1999*, Lugano, Switzerland, July 2003.

[13] P. Ramanathan, E. Steinbach, P. Eisert, and B. Girod, "Geometry refinement for light field compression," in *Proc. IEEE Intl. Conf. on Image Processing ICIP-2002*, vol. 2, Rochester, NY, USA, September 2002, pp. 225–228.

[14] C.-L. Chang, X. Zhu, P. Ramanathan, and B. Girod, "Shape adaptation for light field compression," in *Proc. IEEE Intl. Conf. on Image Processing ICIP-2003*, Barcelona, Spain, September 2003.

[15] P. Ramanathan, M. Kalman, and B. Girod, "Rate-distortion optimized streaming of compressed light fields," in *Proc. IEEE Intl. Conf. on Image Processing ICIP-2003*, vol. 3, Barcelona, Spain, September 2003, pp. 277–280.

[16] P. Ramanathan and B. Girod, "Theoretical analysis of geometry inaccuracy for light field compression," in *Proc. IEEE Intl. Conf. on Image Processing ICIP-2002*, vol. 2, Rochester, NY, USA, September 2002, pp. 229–232.

[17] R. Gallager, *Information Theory and Reliable Communication*. Wiley, 1968.