Rate-distortion Optimized Streaming of Compressed Light Fields with Multiple Representations

Prashant Ramanathan and Bernd Girod
Department of Electrical Engineering
Stanford University
Stanford CA 94305

***Abstract*— Image-based rendering has been proposed as a way of enabling interactive photorealistic viewing of objects and scenes without the complexity of traditional computer graphics rendering techniques. Image-based rendering, however, relies on a large amount of image data to achieve photorealistic quality and freedom in viewing directions and position. This poses several challenges for remote viewing of these data sets over a network, the first of which is efficient compression of the image data. When streaming this data to a remote user who is interactively viewing the light field, random access to images also turns out to be an important consideration. With conventional light field coding techniques, there is typically a trade-off between random access and compression efficiency. Recently, a new encoding scheme using multiple representations based on SP-frames from video coding has been proposed. This new scheme provides both good random access and compression efficiency. In this paper, we propose a method for doing rate-distortion optimized streaming of light fields that have been encoded with multiple representations. We demonstrate that using multiple representations in a streaming scenario can provide better rate-distortion performance to the remote user.**

I. Introduction

3-D content is now commonplace on the Internet, but most content is limited to data sets with few images and limited mobility around the scene or object. Remote interactive viewing of high-quality, photo-realistic 3-D objects and scenes can enable new applications in virtual reality, gaming, virtual museums and e-commerce.

Image-based rendering has been proposed for interactive applications as an alternative to traditional graphics rendering techniques that are computationally complex. Novel views can be computed by simply re-sampling acquired image data, and using a geometry model if available. In this work, we consider a image-based rendering data set called a *light field*.

A light field [1], [2] is an image data set, that represents the outgoing radiance from a particular scene or object, at all points in 3-D space and in all directions. This 4-D data set 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.

Because of the large amount of image data, light fields must be efficiently represented. The most efficient compression techniques use disparity compensation, which utilizes geometry information to predict one image from one or more other images. This technique is similar to motion-compensation in video that utilizes motion information to predict one frame from other frames. In our work, we consider a closed-loop prediction-based light field coder that uses disparity compensation with a geometry model [3]–[5].

Most work has looked at the storage or download transmission scenario. An interesting viewing scenario is one where light field images are streamed to a user who interacts with the 3-D object or scene. The key problem here is to select which images to send or re-transmit, based on what the user is looking at, the network conditions, knowledges of what has already been sent and received, and the importance of each image. In [6], we describe a rate-distortion optimized packet scheduling framework for the streaming of compressed light fields, that attempts to maximize the rendered image quality for the user with a rate constraint.

The prediction structure used for encoding light fields can significantly affect the streaming performance. Conventional light field compression algorithms use prediction to reduce the size of each image, but with an increases in the random access cost. For streaming, random access to images is critical, as will be shown in Section III. In [7], we propose a new encoding method using multiple representations that incorporates random access capabilities and predictive coding efficiency. In the current paper, we propose a method for doing rate-distortion optimized packet scheduling for light fields that are encoded with multiple representations.

The outline of the paper is as follows. In Section II, we review the framework for rate-distortion optimized streaming of light fields. We compare the performance of streaming independently encoded light field images versus a hierarchical predictive encoding of the images in Section III. In Section IV, we describe our approach for encoding light fields images using multiple representations. We propose a method for rate-distortion optimized streaming of light fields with multiple representations in Section V. Finally, in Section VI, we present our experimental results.

II. Rate-distortion optimized streaming of light fields

In this section, we review our framework for rate-distortion optimized streaming of compressed light fields [6]. This work is based on earlier work on packet scheduling for audio and video streaming [8], [9].

Our light field coder uses prediction between images to encode the light field images. A data unit considered for

transmission to the remote user contains the information for a particular image. If that image is predicted from other images, then the data unit is dependent on other data units to be correctly decoded. In [8], [9], this dependency is captured by an acylic directed graph.

The goal in packet scheduling is to determine at what times to send and potentially retransmit a data unit in order to minimize the distortion that the remote user experiences given some rate constraint over the network. Mathematically, we can describe this as minimizing the function

$$J(\boldsymbol{\pi}) = D(\boldsymbol{\pi}) + \lambda R(\boldsymbol{\pi}), \tag{1}$$

where J is the Lagrangian cost, D is the distortion the user experiences, R is the rate over the network, and λ is the Lagragian trade-off parameter. $\boldsymbol{\pi}$ represents the transmission policy, which indicates the schedule of transmissions of each data unit. We assume that these transmissions occur at fixed intervals, such every 100ms, as in our case.

The challenge in performing this minimization consists of estimating the distortion D given our policy and the network conditions, and optimizing over the large parameter space of policies $\boldsymbol{\pi}$. The latter problem is addressed in [8], [9] by iterative minimization over the policies of each data unit in turn, while holding the policies of the other data units constant.

Estimating the distortion poses special challenges for light field streaming, detailed in [6]. Most obvious is that the distortion depends on the user's viewing trajectory. The importance of a particular image depends upon where the user is currently looking. We take the viewing trajectory into account explictly when calculating the distortion contribution for a set of data units.

In [8], [9], an additive distortion model was used. For a light field, the importance of a particular image highly depends on whether or not highly-correlated neighboring images are available at the receiver. We show in [6] how to consider various combinations of images in our distortion calculation in a tractable manner.

Finally, with light field rendering, an image or data unit may be required for more than one view. Thus, a data unit may be required at several time instances, corresponding to the views for which it is needed. The time instance by which a data unit must arrive to be decoded and rendered is called the decoding deadline. In [6], we generalize the original rate-distortion optimized streaming framework to deal with the multiple deadlines that exist in light field streaming.

We showed in [6] that the rate-distortion optimized streaming framework can provide signficantly better performance over a simple heuristic streaming approach. In the next section, we see the effect of the encoding prediction structure on the streaming performance.

III. Random access and streaming performance

In this section, we compare the streaming performance for two different light field encoding prediction dependency structures. The first prediction structure "INTRA" uses independently encoded images with no prediction between images. This prediction structure has very good random access to images, but the compressed size of each image is typically larger than when using prediction between images. The second scheme uses a hierarchical prediction structure, where images are categorized into different levels. "Level 0" images are independently encoded, "Level 1" images are predicted from the nearest "Level 0" images, and so on. Here the size of each compressed image is smaller, but random access to the images is limited because several images may need to be transmitted and decoded to decode a particular image.

We perform rate-distortion optimized streaming using two light field data sets. The *Bust* light field contains 339 images, each of resolution 768×480, and the *Horse* data set contains 110 images, each of resolution 512×512. Both light field use a hemispherical camera arrangement, and have corresponding geometry models to be used for rendering and compression.

We perform rate-distortion optimized streaming for various viewing trajectories of the two data sets. In the following results, we consider a set of 10 random viewing trajectories, each consisting of 25 views. In these random trajectories, since the views are spaced close together, we call these trajectories *dense*.

We assume a network that independently loses data units with a probability of 0.1%. With data units that it does not lose, it delays them according to a gamma distribution with a mean of 50ms and standard deviation of 23ms.

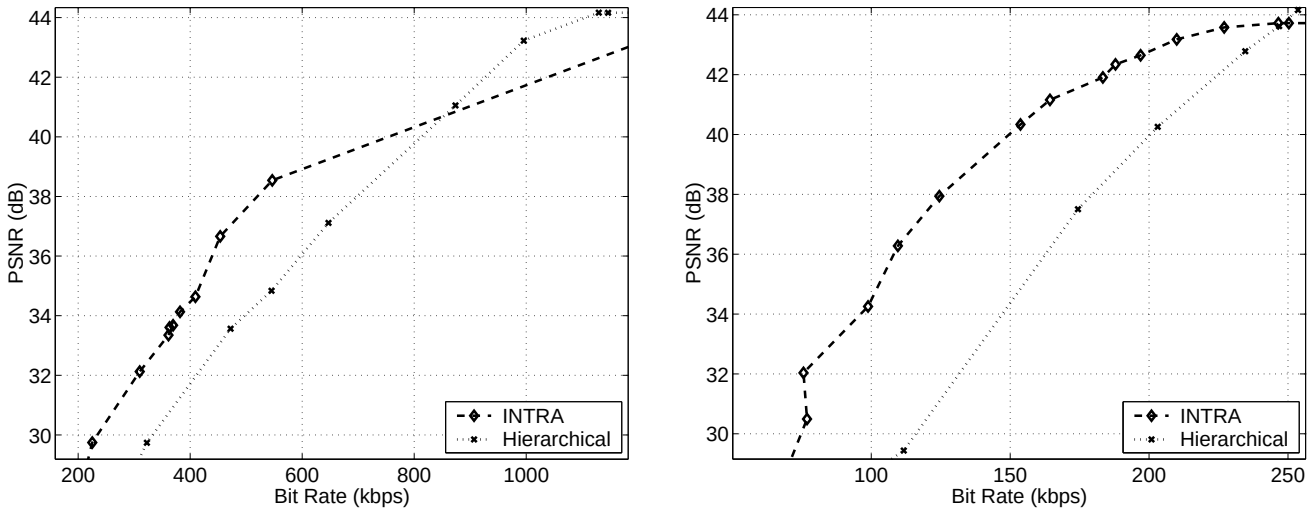
In our streaming system, we have some latency between the user requesting a view and when it is rendered on their screen. We assume that the server has 200ms between when it knows a view is needed, and when the images for that view must arrive at the remote client. The view trajectory could be transmitted from the user to the server and known perfectly, adding additional latency to the system, or could be predicted from past user behavior. We assume that we know the desired view perfectly at the server 200ms before its decoding deadline.

In our system, we consider transmissions every 100ms. The 200ms delay provides the server 2 transmission opportunities in which to send the related images. Described another way, for a given transmission opportunity, the server considers 2 views when computing transmission policies. We call this our *view window*.

Figure 1 compares the rate-distortion optimized streaming performance of the INTRA and hierarchical coding for the *Bust* and *Horse* data sets, and the *dense* viewing trajectories. The results are averaged over the 10 random trajectories.

Consider the results for the *Bust* data set, shown in Figure 1(a). The x-axis shows the transmitted bit-rate in kbps, and the y-axis represents the image quality in terms of dB of PSNR. We see that for a large range of bit-rates, independent coding outperforms hierarchical prediction. We see even more impressive differences for the *Horse* light field, shown in Figure 1(b).

When encoding the entire light field data set, hierarchical prediction is a much more efficient strategy for encoding the data set. Despite that fact, in a streaming scenario, for these



(a) *Bust* (b) *Horse*

Fig. 1. Rate-distortion optimized streaming results for the *Bust* and *Horse* light fields, averaged over 10 *dense* trajectories. INTRA coding results in streaming performance that is better than or is comparable to Hierarchical coding.

data sets, it appears that random access is as important or more important, and independent coding is the more efficient strategy. In next section, we will describe an encoding scheme with the same random access capabilties of independent coding, but using prediction to give superior compression performance.

IV. Multiple Representations Encoding

Conventional coding of light field images uses a fixed prediction structure where one image is independently coded or predicted from other images. With this fixed structure, in order to communicate an image that is predicted, the images used for prediction must also be transmitted to the decoder.

An alternative to this would be to predict based on what is already available at the decoder. For different combinations of prediction images available at the decoder, we could have different image representations that we would store at the server and send as appropriate. The problem with this *multiple representations* approach using conventional image and video coding techniques is that, due to quantization, different representations will lead to different reconstructed images. This prediction mismatch can propagate during subsequent prediction steps.

In [10], the authors solve the mismatch problem using coset codes. We present a simpler solution to this problem in [7], where we eliminate prediction mismatch by using standard video coding concepts. Our approach, in addition, does not have the decoding complexity of coset codes. We summarize our work in this section.

In conventional predictive coding, the prediction signal is subtracted from the original image giving the residual signal, which is then transformed, quantized and entropy coded. Due to the quantization of the transformed residual image, the reconstructed image obtained for different predictions are not, in general, identical.

SP-frames have been used in video coding to provide identical image reconstruction for different prediction signals [11]. Figure 2 shows a diagram of the encoder and decoder that we implement, based on SP-frames. In the encoder, the original image is first transformed and quantized, as is the prediction image. These two sets of quantization indices are then subtracted from one another to give a set of indices that are entropy coded.

The decoder reverses this signal path by decoding the bitstream and adding back the quantization indices from the prediction image. Since we have identical prediction signals at both the encoder and decoder, the output of the summation is exactly the quantized coefficients of the original image, that is independent of the prediction signal used. Thus, even if we use different prediction images, we obtain identical image reconstructions.

In the context of SP-frames, the approach we have just described represents secondary SP-frames. We also use an INTRA coded image representation which we consider our primary SP-frame. The INTRA coded image uses a quantizer which is identical to that used for all the other representations for the image.

In our light field coding application, for each image we consider K different prediction images. These, for instance, could be the prediction image from each of the K neighbors of this image. Or, they could include combinations of images to be used for prediction. In total we have $K+1$ representations, corresponding to the K prediction images, and the INTRA representation of that image. The image reconstructions from the K prediction signals will be identical to that of the INTRA

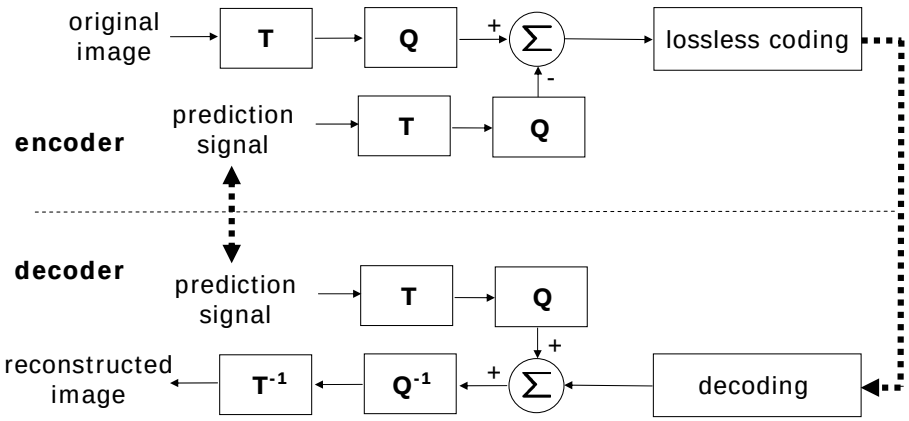


Fig. 2. Encoding and decoding of images with no prediction mismatch Regardless of the prediction signal used, the reconstructed image will be identical to that of INTRA coding. This system is equivalent to secondary SP-frames with INTRA-coded primary SP-frames.

reconstructed image.

The representations for each image are pre-encoded and stored on the server. During interactive streaming, the appropriate image representation must be sent to the remote viewer. In the next section, we propose a method of selecting in a rate-distortion optimized fashion, the image representations to be sent.

V. Streaming Multiple Representations

When streaming with multiple representations, we now have a large number of data units to be considered for transmission. If we have a large number of representations for each image, this can quickly become computationally inefficient or even intractable.

We propose a simple two-step procedure. First, we use rate-distortion optimized packet scheduling using only the INTRA coded representations to select the images. Second, for the selected image, we transmit the representation with the lowest bit-rate that can be decoded.

We start by scheduling using only the INTRA representations since there are no prediction dependencies with this scheduling, and this is a special case of streaming with hierarchical prediction encoded images. Since we know that each representation of an image will result in identical image reconstructions, the estimated distortion using the other representations will be the same as that for the INTRA representation. The different rates of the other representations are not taken into account, however, which may result in a suboptimal solution.

Once we know which images to send, we then select the appropriate representation for each image. We do this in a conservative manner by only considering representations that we know can be decoded at the receiver. We know which images have been acknowledged, and therefore which images can be used for prediction, and which representations are known to be decodable.

Of these representations, which always includes the INTRA representation, we select the one with the smallest rate. Using such a scheme guarantees that we do no worse than INTRA coding since we have exactly the same image reconstruction with no greater rate. The next section details the experimental results from using multiple representations for streaming.

VI. Experimental Results

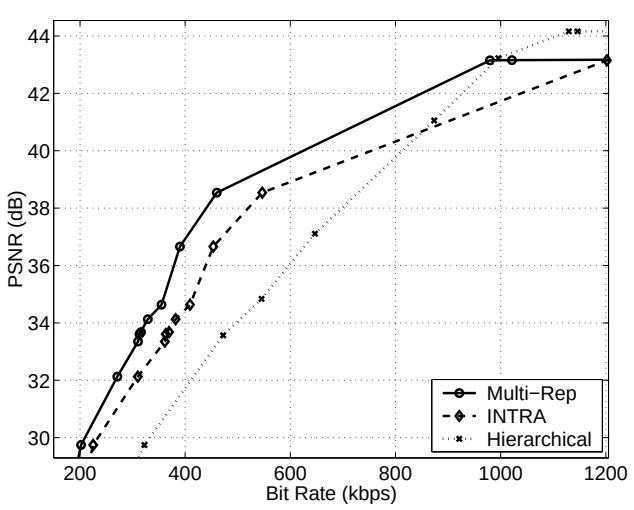
We consider the same experimental set-up as before, with the *Bust* and *Horse* data sets. We perform rate-distortion optimized streaming for various viewing trajectories of the two data sets. For each data set, we consider two sets of 10 random viewing trajectories, each consisting of 25 views. The first set of trajectories, which we name *dense* consists of views that are closely spaced relative to one another. The second set of trajectories, which we name *wide* consists of views that are widely spaced relative to each other. These two types of trajectories can be considered the two extremes of viewing behaviour: slowly examining an object by rotating it, or moving around the object rapidly to get a quick overall impression of it. The network parameters are also identical to those before. We use $K = 31$ neighboring images, or a total of 32 representations for each image.

We compare the rate-distortion streaming performance using multiple representations versus INTRA coding and hierarchical coding in Figures 3 and 4. Figure 3(a) shows the comparison for the *Bust* light field and *dense* trajectories. Here, we see that streaming using multiple representations encoding can improve image quality by up to 2 dB over INTRA coding, or reduce the bit-rate by 15%. For *wide* trajectories, shown in Figure 3(b), we see smaller improvements since we have to predict from more distant images with less correlation.

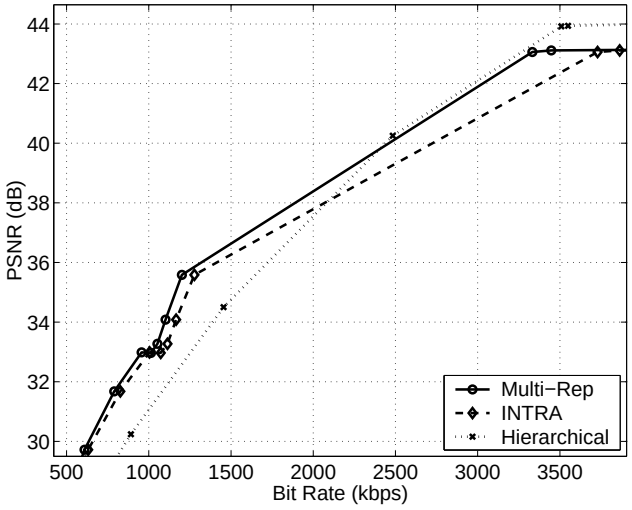
Figures 4(a) and 4(b) show the streaming performance for the *Horse* light field, with similar results as that of *Bust*. We see an improvement of up to 2 dB in image quality at the same bit-rate, or a reduction of up to 10% in bit-rate at the same image quality. Larger improvements are seen for the *dense* trajectories than for the *wide* trajectories.

VII. Conclusion

For interactive streaming of light fields, random access to images is critical to the rate-distortion performance of the system. A recently proposed *multiple representations* light

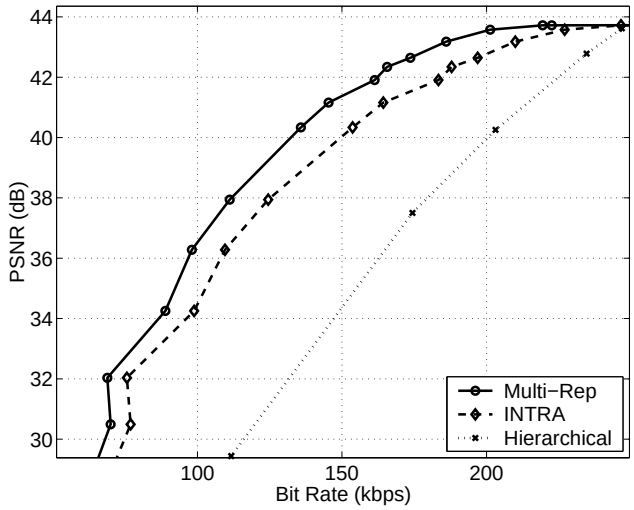


(a) *dense* trajectories

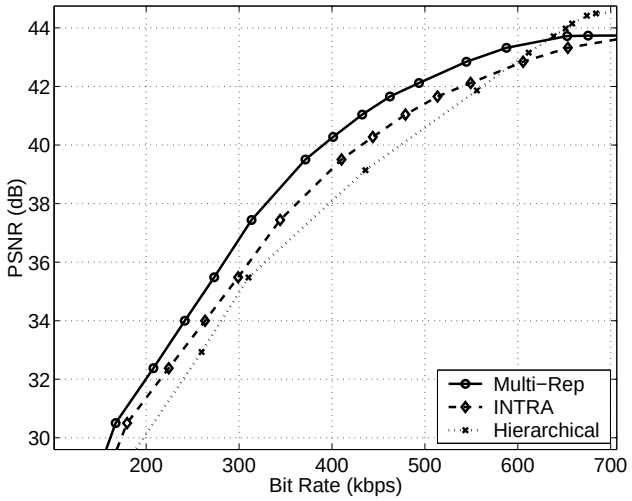


(b) *wide* trajectories

Fig. 3. Rate-distortion optimized streaming results for the *Bust* light field, averaged over 10 trajectories. Multiple representations coding has superior streaming performance compared to INTRA coding. There are improvements of up to 2dB in image quality at the same bit-rate, or a bit-rate reduction of up to 15% at the same image quality.



(a) *dense* trajectories



(b) *wide* trajectories

Fig. 4. Rate-distortion optimized streaming results for the *Horse* light field, averaged over 10 trajectories. Multiple representations coding has superior streaming performance compared to INTRA coding. There are improvements of up to 1.5dB in image quality at the same bit-rate, or a bit-rate reduction of up to 10% at the same image quality.

field encoding scheme provides random access to images while more efficiently encoding the light by using prediction. We presented a method for rate-distortion optimized streaming of a light field encoded using multiple representations. This method used a simple two-step approach where the image to be transmitted is first selected, then the appropriate representation of that image chosen. Using this approach, we showed superior performance over INTRA coded light field images. The image quality, in terms of PSNR, was improved by up to $1-2$ dB for the same bit-rate, or the bit-rate was reduced by up to $10-15\%$, at the same image quality, for the data sets and trajectories we studied.

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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] P. Ramanathan and B. Girod, "Random access for compressed light fields using multiple representations," in *Multimedia Signal Processing Workshop MMSP-2004*, 2004, accepted.

[8] P. A. Chou and A. Seghal, "Rate-distortion optimized receiver-driven streaming over best-effort networks," in *Packet Video Workshop*, Pittsburgh, PA, USA, April 2002.

[9] P. A. Chou and Z. Miao, "Rate-distortion optimized streaming of packetized media," Microsoft Research, Tech. Rep. MSR-TR-2001-35, February 2001, (also submitted to IEEE Transactions on Multimedia).

[10] A. Jagmohan, A. Seghal, and N. Ahuja, "A state-free causal video encoding paradigm," in *Proc. ASILOMAR Conf. on Signals and Systems 2003*, Pacific Grove, CA, November 2003.

[11] M. Karczewizcz and R. Kurceren, "The SP- and SI-frames design for H.264/AVC," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 637–644, July 2003.