THE PLENOPTIC VIDEOS: CAPTURING, RENDERING AND COMPRESSION

Shing-Chow Chan, King-To Ng, Zhi-Feng Gan, Kin-Lok Chan, and Heung-Yeung Shum**

Department of Electrical and Electronic Engineering
The University of Hong Kong
{ scchan, ktng, zfgan, kinlok }@eee.hku.hk

* Microsoft Research Asia
hshum@microsoft.com

ABSTRACT

This paper presents a system for capturing and rendering a dynamic image-based representation called the plenoptic videos. It is a simplified version of light fields for dynamic environment, where user viewpoints are constrained along the camera plane of a linear array of video cameras. The system consists of a camera array of 8 Sony CCX-Z11 CCD cameras and eight Pentium 4 1.8 GHz computers connected together through a 100 BaseT LAN. Important issues such as multiple camera calibration, real-time compression, decompression and rendering are addressed. Experimental results demonstrated the usefulness of the proposed parallel processing based system in capturing and rendering high quality dynamic image-based representation using off-the-shelf equipment, and its potential applications in visualization and immersive television systems.

1. INTRODUCTION

Image-based rendering (IBR) has recently emerged as a promising alternative to 3D computer graphics for photo-realistic rendering of scenes and objects from a collection of densely sampled images. Central to IBR is the plenoptic function [1], which forms a new framework for developing sophisticated virtual reality and visualization systems. Another important advantage of IBR is the superior image quality that it offers over 3D model building, especially for very complicated real world scenes. It also requires much less computational power for rendering, regardless of the scene complexity. Unfortunately, image-based representations usually consist of hundreds or thousands of images, which involve large amount of data. To simplify the capturing and storage of the plenoptic function, various image-based representations with lower dimensions have been advocated [2]-[5]. Most image-based representations reported so far deal with static scenes. This is largely attributed to the logistical difficulties in capturing and transmitting dynamic representations, which involve huge amount of data. In fact, it has simulated considerable research effort into efficient compression methods for various image-based representations such as the light field, lumigraph, and concentric mosaics [5]-[11]. A study of real-time capturing, compression and rendering of image-based representations for dynamic environment is thus highly desirable.

Towards this goal, we construct a system for real-time capturing, compression and rendering of a simplified light field for dynamic scenes. We coined these simplified dynamic light fields (SDLF) [18] the plenoptic videos, because of their close relationship with traditional videos supporting multiple viewpoints. Through this system, it was also demonstrated that how parallel processing and inexpensive equipment can be utilized to capture and process the image-based representation of dynamic scenes efficiently and mostly in real-time, which is one of the major obstacle in dynamic IBR research. Unlike capturing static image-based representations, methods for calibrating multiple cameras and compressing the plenoptic videos have to be developed.

One immediate application of the proposed system is "interactive 3D electronic catalog or brochure" and "short advertisement plenoptic video clip," where the plenoptic videos are distributed either in form of DVDs or the Internet for viewing by potential customers in a computer (mouse driven). This was illustrated in our demos, consisting of a glass music box and two lead crystals, which are usually very difficult to model with photo-realistic quality. Another possible application is a head and shoulder-type 3D videophone, where the depth variation, like the music box sequence, is relatively small. The paper is organized as follows. The proposed plenoptic video is described in Section 2. Then, the developed capturing system is detailed in Section 3. Section 4 is devoted to the rendering and compression of the plenoptic video. The experimental results are given in Section 5. Finally, conclusions are drawn in Section 6.

2. PLENOPTIC VIDEOS

Previous attempts to generalize image-based representations to dynamic scenes are mostly based on 2D panoramas. These include the QuickTime VR [2] and panoramic videos [10]. The panoramic video is a sequence of panoramas created at different locations along a path in space, which can be used to capture dynamic scenes at a stationary location or in general along a path with 360 degrees of viewing freedom. The plenoptic video described in this paper is a simplified light field for dynamic environment, where the viewpoints of the user are constrained along a line instead of a 2D plane in [4]. This greatly reduces the complexity of the dynamic IBR system. However, unlike panoramic videos, users can still observe significant parallax and lighting changes along the horizontal direction.

More recently, there were attempts to construct light field video systems for different applications and characteristics. These include the Stanford multi-camera array [12], the 3D rendering system of Naemura *et al*. [13], and the (8×8) light field camera of Yang *et al*. [14]. The Stanford array consists of more than one hundred of cameras and is intended for large environment applications. It uses low cost CMOS sensor and dedicated hardware for real-time compression. The systems in [14], [15] consist of respectively 16 and 64 cameras and are intended for real-time rendering applications. Unlike the Stanford array and our system, they do not support real-time compression of the captured videos. Our system chooses different design tradeoffs, which yields very good rendering quality. Our intention is not to compare directly with these systems, but rather to disseminate useful system construction experience and compression techniques for other researchers to build inexpensive arrays and fairly high quality rendering systems using off-the-shelf equipment. The latter has been one of the major obstacles in the study of dynamic IBR.

3. THE PLEOPTIC VIDEO SYSTEM

Fig. 1 shows the block diagram of our plenoptic video capturing and processing system. A set of synchronized video cameras is used to capture the light field images at each time instant to form sequences of videos. The video signals are then fed to the real-time video compression boards in the parallel processing

system, which consists of a number of PCs connected together through a high-speed network such as the 100 BaseT or gigabit Ethernet. With the advent of video compression hardware, inexpensive real-time MPEG-2 compression boards are now readily available. The compressed videos will be stored directly to the hard disk of the PCs. Again thanks to the advent of PC technology, high-speed and inexpensive hard disks with 120 GB of storage are now in common use. As a result, the parallel arrangement allows us to capture 4D dynamic light fields for a fairly long period of time, say several hours.

To avoid unnecessary complication, our prototype system employs 8 (up to 10) cameras as a reasonable tradeoff between hardware complexity and performance. Our system uses closely spaced CCD cameras to reduce problems due to insufficient sampling [19] and avoid the large variations of the CMOS cameras, which usually complicate camera calibration. Our system is also relatively easy to construct, as it requires only off-the-shelf components and readily available equipment. Another valuable feature of our system is its distributed nature, which allows us to capture, compress, process, and render the plenoptic video efficiently. The real-time rendering of the plenoptic video will be described in more details in Section 4.1. We believe that parallel processing is essential to handle the demanding storage and computational requirements of plenoptic videos and other dynamic image-based representations. Although our prototype system has a linear configuration, other similar configuration such as (2×8) or (3×8) are possible and it will improve the viewing freedom of the users. We believe that the resulting rendering quality will be similar to the one reported here. Due to the use of data compression and parallel processing, our system is reasonably scalable.

3.1 Construction

In our prototype system, eight Pentium 4 1.8 GHz computers are connected together through a 100 BaseT LAN. At 1.8 GHz, it is possible to perform software-assisted real-time MPEG-2 compression at a resolution of (720×480) using the Pinnacle PCTV capturing board. A camera array using 8 Sony CCX-Z11 CCD cameras is constructed as shown in Fig. 2. The outputs are in NTSC format (525-line interlaced video at 25 f/s) and they are synchronized by modifying the electronics inside the cameras so that they operate on the same clock signal. The spacing between successive cameras is 2.5 cm and four tuning screws are used to control the tilting angles of each camera. Note, all these components are relative inexpensive and they can readily be extended to include more cameras.

3.2 Camera Calibration

During construction, the camera lenses are carefully installed to the hardware stand and similar focuses and tilting angles are maintained. The cameras are then calibrated using the method in [16]. This method is originally proposed for calibrating a single camera and the relative position of the camera and the viewing angle with respect to a reference grid position can be estimated. More precisely, five images (the grid images) of a certain grid pattern, which consists of squares evenly spaced at a regular grid, are taken by the camera at five different positions. The corners of the squares in each grid image are then determined in order to recover the intrinsic and extrinsic parameters of the cameras. This information allows us to correct the geometric distortion of the camera lens, determine the relative positions and viewing angles of the cameras with respect to the reference grid. After a reference camera plane is chosen, the images captured from all the other cameras can be wrapped to the same coordinate of the reference camera. The rectified videos of the cameras are more amenable to rendering.

4. RENDERING AND COMPRESSION OF PLENOPTIC VIDEOS

4.1. Selective Transmission and Rendering

If the plenoptic video is decoded into raw images and stored in a hard disk, real-time rendering can be readily achieved. However, the memory requirement is very large and the playback time is limited. If the plenoptic video is rendered from the compressed bit stream, then even with the use of selective decoding the computational requirement for the decoding and rendering is very large. The basic idea of selective transmission/rendering is to decode in parallel the multiple streams of the videos in a network of computers, and transmit those pixels required to the rendering machine over the network, possibly with simple compression. This offloads the rendering machine at the expense of longer user response time. However, we believe that selective transmission is essential to the distribution of plenoptic video in future applications.

4.2. Depth Estimation Problem

For our captured plenoptic video sequences, the depth variation is relatively small and according to the plenoptic sampling analysis [19], the rendering artifacts will be insignificant as long as the focus plane is chosen as the mean depth of the scene. This was verified in our synthetic as well as real plenoptic video experiments. For complicated scenes, more geometry information such as the depth map in lumigraph and the geometric model in the surface light fields would be required to avoid the rendering artifacts due to incorrect depth values.

4.3. Requirements for IBR Compression

The plenoptic video is a densely sampled higher dimensional signal. Their data sizes are huge but their samples are highly correlated. Direct application of traditional compression algorithms, however, usually results in sub-optimal performance. Providing *random access* to the entropy coded data such as Huffman or arithmetic coding for real-time rendering is another important problem. With efficient random access mechanism, *selective decoding* [9], [10] or just-in-time (JIT) decoding [6], [11] can be employed to decode on-line those pixels which are required for rendering. Similar problems exist in the transmission of plenoptic videos and techniques to support their *selective transmission/reception* are of paramount importance.

4.4. Temporal and Disparity Compensation

The compression of plenoptic videos is closely linked to conventional video compression. However, as video streams in a plenoptic video are taken at nearby positions in a 1D array, they appear to be shifted relative to each other, because of the disparity of image pixels. Spatial prediction or disparity compensated prediction has been used in coding of static light fields [6]-[8] and stereo image coding [17]. The coding algorithm considered here can be viewed as their generalization to the dynamic situation. An MPEG-2 like algorithm for coding the plenoptic video was proposed in [18]. Multiple videos adjacent to each other are compressed together to form a group of field (GOF). There are two types of video streams in a GOF: *main* and *secondary* video streams. Main video streams are encoded using the MPEG-2 algorithm, which can be decoded without reference to other video streams. The *I*-pictures from the secondary stream are encoded using *spatial prediction* from the reference *I*-picture in the main stream. It is because adjacent images appear to be shifted relative to each other, similar to the effect of linear motion in video coding. *P*- and *B*-pictures from the secondary stream can be predicted using spatial prediction from adjacent *P*-picture in the main stream or the motion compensation from the reference *I*- or *P*-pictures in the same secondary stream. To provide *random access to individual pictures*, we have adopted a modified MPEG-2 video compression algorithm [10] to encode the image frames.

5. EXPERIMENTAL RESULTS

Two plenoptic videos are captured by the system: *Glass Music Box* and *Crystal Dragon*. In the *Glass Music Box*, a glass music box was placed at the center of the scene and it was rotating at a regular speed. A moving spotlight was used to change dynamically the lighting of the scene. It can be seen from the images that significant lighting changes, reflections, and parallax were captured. In addition, it resembles closely a traditional video except that the viewpoint can be continuously changed along the line containing the cameras. Fig. 3 shows a snapshot of another plenoptic video *Crystal Dragon* captured by the system (rectified). The sequence consists of a lead crystal in the shape of a dragon, which was placed on a wooden platform. Beside it is another crystal turtle, which was placed on a lighting platform that changes color periodically. A lighting candle and a moving spotlight were also included to demonstrate the lighting changes and reflective properties of the scene. Since the capturing system is able to handle videos of more than an hour, the two plenoptic videos were taken in a single shot. The distances between the objects and the camera array were also varied to evaluate the effect of camera calibration to the rendering quality. Each uncompressed video stream consumes about 30 GB of storage and the entire plenoptic video requires 240 GB of storage.

5.1. Rendering

Using selective transmission, we are able to stream continuously plenoptic video with the resolution of 256×256 at a frame rate of 15 f/s over the network. Due to network delay, there is a slight delay in the user response. The frame rate and the resolution can be increased if the raw data stream is compressed by simple coding method such as vector quantization. For rendering from raw data in the hard disk, real-time rendering can be achieved with a resolution of 720×480 and a frame rate of 15 f/s. It can be seen that the lighting changes and reflective properties of the glass and lead crystal were well captured. It was found that slight artifacts are still present in some of the rendered images. These artifacts usually appear as ghost images or blurring of the images, depending on the values of the mean depth and the quality of camera calibration. In addition, it was found that the artifacts are less noticeable if the objects are farther away from the camera planes. It is because of the reduced resolution of the images as well as the reduced sensitivity of the image pixels to errors in camera positions.

5.2. Compression

The plenoptic videos consist of 8×1 24-bit RGB videos with 720×480 pixels. Compression results for different number of video streams in a GOF are investigated and plotted in Fig. 5. For SP3, SP5 and SP7, we have three, five and seven video streams in the GOF, respectively [18]. As a comparison, we also compressed all the video streams of the plenoptic videos by MPEG-2 algorithm independently. It can be seen that the performances of the proposed algorithm using both temporal and disparity compensation has significant improvement over the independent coding scheme. This shows that there is significant amount of spatial redundancy among the video sequences. When the number of video streams in the GOF is increased, the PSNR improves because fewer *I*-pictures are coded and better disparity prediction is obtained in the plenoptic video. However, the difference between SP5 and SP7 is small because disparity compensation will be less effective when video streams are far apart. In Fig. 4, the typical reconstructed images show good quality of reconstruction at 583 kbps per stream and compression ratio (C. R.) of 341 for *Glass Music Box* and 624 kbps and C. R. of 319 for *Crystal Dragon*.

6. CONCLUSION

We have presented a novel system for capturing, compression and rendering of simplified dynamic light fields, called the plenoptic videos. By appropriate system design, we have demonstrated that dynamic image-based representations of high dimensionality can be captured and processed using off-the-shelf components and readily available equipments. Methods for calibrating multiple cameras and compressing video data in the plenoptic video system were also developed. One application of the proposed plenoptic video system is in "interactive 3D electronic catalog or brochure" and "short advertisement plenoptic video clip," where the plenoptic videos are distributed either in form of DVDs or the Internet for viewing by potential customers in a computer (mouse driven). This was demonstrated in our demos, consisting of a glass music box and two lead crystals, which are usually very difficult to model with photo-realistic quality. Finally, we hope the experience and findings in this work will facilitate the further development and widespread use of dynamic image-based representations as an efficient means for visualization, especially for 3D immersive TV systems.

REFERENCES

[1] E. H. Adelson and J. Bergen, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing*, pp. 3-20, MIT Press, Cambridge, MA, 1991.

[2] S. E. Chen, "QuickTime VR–an image-based approach to virtual environment navigation," in *Proc. of SIGGRAPH*, pp. 29-38, Aug. 1995.

[3] S. J. Gortler, R. Grzeszczuk, R. Szeliski and M. F. Cohen, "The lumigraph," in *Proc. of SIGGRAPH*, pp. 43-54, Aug. 1996.

[4] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. of SIGGRAPH*, pp. 31-42, Aug. 1996.

[5] H. Y. Shum, S. B. Kang and S. C. Chan, "Survey of Image-based Representations and Compression Techniques," to appear *IEEE Trans. Circuits and Systems for video technology,* 2003.

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

[7] M. Magnor and B. Girod, "Data compression for light-field rendering," *IEEE Trans. on CSVT*, vol.10, no.3, pp. 338-343, Apr. 2000.

[8] X. Tong, et al., "Coding of multi-view images for immersive viewing," in *Proc. of IEEE ICASSP,* vol. 4, pp. 1879-1882, June 2000.

[9] H. Y. Shum, K. T. Ng and S. C. Chan, "Virtual reality using the concentric mosaic: construction, rendering and data compression," in *Proc. of IEEE ICIP*, vol. 3, pp. 644-647, Sept. 2000.

[10] K. T. Ng, S. C. Chan, H. Y. Shum and S. B. Kong, "On the data compression and transmission aspects of panoramic video," in *Proc. of IEEE ICIP,* vol. 2, pp. 105-108, Oct. 2001.

[11] J. Li, H. Y. Shum and Y. Q. Zhang, "On the compression of image based rendering scene: a comparison among block, reference and wavelet coders," *International Journal on Image and Graphics*, vol. 1, no. 1, pp. 45-61, 2001.

[12] B. Wilburn et al., "The Light Field Video Camera," in *SPIE Proc. Electronic Imaging: Media Processors*, vol. 4674, Jan. 2002.

[13] T. Naemura, J. Tago and H. Harashima, "Real-time video-based modeling and rendering of 3D scenes," *IEEE Computer Graphics and Applications*, pp. 66-73, Mar.-Apr. 2002.

[14] J. C. Yang, M. Everett, C. Buehler and L. McMillan, "A real-time distributed light field camera," in *Proc. of Eurographics Workshop on Rendering*, pp. 77-86, 2002.

[15] B.Goldl¨ucke, M. Magnor and B. Wilburn, "Hardware-accelerated Dynamic Light Field Rendering," in *Proc. of VMV*, pp.455-462, 2002.

[16] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. on PAMI*, vol. 22(11), pp. 1330-1334, 2000.

[17] W. Woo and A. Ortega, "Stereo image compression based on the disparity compensation using the MRF model," in *Proc. of SPIE Conference on VCIP*, vol. 2727, pp. 28-41, 1996.

[18] S. C. Chan, K. T. Ng, Z. F. Gan, K. L. Chan, and H. Y. Shum, "The compression of simplified dynamic light fields," in *Proc. of IEEE ICASSP*, vol. 3, pp. 653-656, Apr. 2003.

[19] J. X. Chai, X. Tong, S. C. Chan, and H. Y. Shum, "Plenoptic sampling," in *Proc. of SIGGRAPH*, pp. 307-318, July 2000.

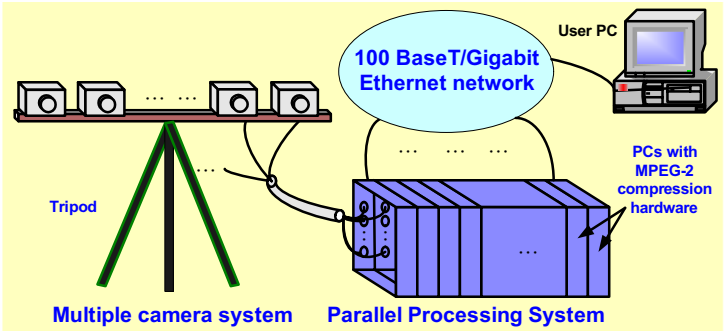

Fig. 1. Block diagram of the plenoptic video system.

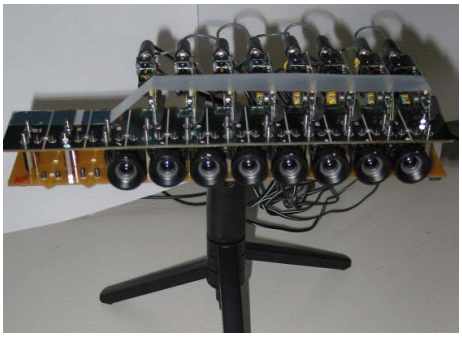

Fig. 2. The 8-camera array of the plenoptic video system.

Fig. 3. Snapshots of the plenoptic video *Crystal Dragon*. Each row consists of the eight images taken from the cameras (form left to right) at a given time instant.

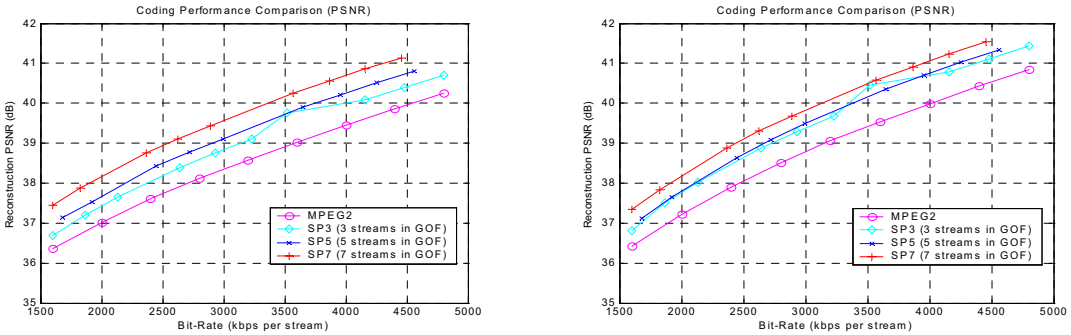

Fig. 5. Coding results of the plenoptic videos. (Left) The *Glass Music Box* sequence. (Right) The *Crystal Dragon* sequence.

Fig. 4. Typical reconstructed images of (Upper) the *Glass Music Box* sequence (583 kbps per stream) and (Lower) the *Crystal Dragon* sequence (624 kbps per stream) in the main (left) and secondary (right) video streams.