A REAL TIME INTERACTIVE DYNAMIC LIGHT FIELD TRANSMISSION SYSTEM

Yebin Liu, Qionghai Dai (Senior Member, IEEE), Wenli Xu

Broadband Networks & Digital Media Lab, Tsinghua University, Beijing 100086, China

ABSTRACT

The ability to interactively and seamlessly roam in the scenario while watching a video through IP network is an exciting visual experience. In this work, we implemented a 3D TV system with real-time data acquisition, compression, internet transmission, light field rendering, and free-viewpoint control of dynamic scenes. Our system consists of an 8×8 light field camera array, 16 producer PCs, a streaming server system and several clients. Multiple video streams are coded in a real time manner that each client can freely selects the streams for novel view rendering. Also, our system minimize the per-user transmission bit rate while maintaining multi-view simul-switching ability for each user. We believe that this is the first real-time internet streaming system that can simultaneously guarantee real time free-view point control, data storage and support arbitrary number of users. The average transmission bit rate for end user is lower than 2Mbps which is suitable for the broadband IP network.

1. INTRODUCTION

To interactively and seamlessly roam in a scenario while watching a streaming video through IP network is an exciting experience but seems hard to be achieved although great progresses had been made in computer graphics domain and video processing domain for decades. 3D Modeling complexity, rendering quality, data quantity and real time interactivity are all the challenges to the realization of such experience. Thanks to Levoy and Hanrahan [1] and Gortler's [2] publication of light field rendering technique, the challenges of scenario modeling and rendering described above begin to be solved. A light field is a collection of light rays following through space in all directions captured by a multi camera array and recorded as multi-view images. With the advent of light field, new images viewed in any position and any direction can be rendered easily with only a little geometry information or even none geometry information involved. This light field rendering technique circumvents the complicated modeling procedure, simplifies the rendering process and achieves photorealistic rendering effect. Since the publication of this static light field rendering technique, there have been researches investigating time-critical lumigraph renderings by Sloan [3] and dynamic reparameterization of light fields. The discovery of these very fast algorithms dismisses parameterization calculations as a bottleneck to real-time light field rendering and dynamic light field (DLF) rendering, thus overcomes the real time challenge of our dream.

For the technical prevalence and promising features depicted above, light field has been attracting more and more attentions in recent years. An Ad-Hoc Group on 3D Audio and Video (3DAV) was found by MPEG community, and omnidirectional video, free view point video, stereoscopic video and depth video have been discussed in this group. Since light field technique is a natural tool to both free view point video and stereoscopic video, light field and its correlated techniques play an important role in 3DAV. For dense spaced camera system, B.Wilburn [5] has implemented an MPEG2 light field camera array to capture and store DLF. J.C.Yang [6] has developed a light field rendering system that can interactively render 3D scene. Line spaced simple DLF compression and rendering system is also developed in [7]. For stereoscopic service, a real time 3DTV transmission system for autostereoscopic display is developed [8] by using 16 regularly spaced cameras and 16 projectors. For multiview streaming service, a real time interactive multiview system [9] is constructed by using 32 arc spaced cameras.

Despite the above works on light field systems and technologies, there is no system that can serve multiple users as a live broadcast streaming system over internet. The functionalities of such system may include: real time DLF compression, real-time rendering, interactive streaming, arbitrary user support, 2D display and 3D display compatibility, minimized per user transmission bandwidth. In this paper, we develop such a system guarantying all these features.

The rest of the paper is organized as follows. Section 2 outlines the architecture and functionalities of our system. We detail the system modules in section 3. Some experimental results are presented and discussed in section 4. Finally, we conclude our work in section 5.

2. SYSTEM ARCHITECTURE

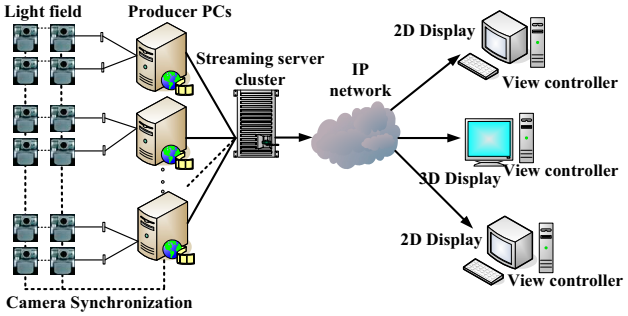

Fig.1. Topology of the real time DLF transmission system

Figure 1 illustrates the topology of a real time DLF transmission system which can be partitioned into three parts: capture part, server part and client part. Geometry calibration, color calibration and data capture belong to the light field capture part. Data compression and data streaming are the works of the server part

This work is supported by the Distinguished Young Scholars of NSFC (No.60525111) and the key project of NSFC (No.60432030)

and the client is responsible for data decoding and new view rendering. Rendering works should be on clients or the burden of the server will increase linearly with the increasing number of the clients if they are finished on the server as [6].

Our system is suitable for both 2D display and 3D display devices since all clients share the same streaming protocol. The server sends the required streams to the clients and the only difference between these clients is the number of new views rendered by the clients. For 2D display, only one new view is rendered while two views are rendered for stereoscopic display and more views for autostereoscopic display. Besides stereoscopic feeling, users can enjoy visual experiences including the following three kinds of view trajectories:

1) View switching: Users are able to change the viewing position and viewing direction as the video continues along time.

2) View zooming: DLF rendering can provide zooming capability for users. If the user trajectories are near the camera plane, a relatively small number of frames are required to generate new views, otherwise, as for the situation of zooming in and out from the camera plane, more frames are needed.

To minimize transmission bandwidth, only the camera-streams required for rendering are streamed to the client. Thus the DLF streaming system involves simultaneously switching of multiple views, which is called simul-switching in this paper.

3. SYSTEM MODULES

In this section, we focus on the major modules including video acquisition, video compression, client display and video streaming.

3.1. Acquisition: Camera calibration and video capture

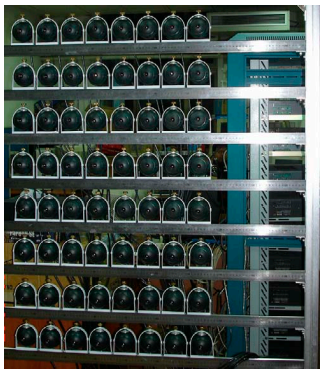

Fig.2. A photo of our 64-camera light field camera array. The cameras are arranged in rows of eight.

Unlike the works in [7] and [9], our cameras are not line located but regularly planar array spaced which provide more freedom and wider visual range for the users. We use 64 BOSER BS-103F color cameras with 320×240, 8 bits per pixel CCD sensors to set up an 8×8 light field camera array (see figure 2). The cameras are connected by IEEE-1394 High Performance Serial Bus to the producer PCs. The maximum frame rate at such resolution is 30 fps. Every 4 cameras (the one camera with its right, bottom and diagonal neighbored cameras) are connected to one of the 16 producer PCs which all have the same hardware configuration: Pentium-IV D CPU 2.8G and 1GB RAM.

A synchronization software is developed to synchronize all the internal clocks of the producer PCs when the system starts up. This procedure is finished in 10ms, thus their clocks differ by no more than 5~10ms. The total time for a producer PC to capture a frame from each of its four connected cameras is no more than 5ms. Therefore, the maximum time differ between any of the two camera in the light filed will be 15ms which may keep the images to within a frame of one another.

The optical axis of each camera is roughly perpendicular to a common camera plane. The horizontal spacing between cameras is about 8cm, and the vertical spacing is about 14cm. A 14×10 checker board are used to calibrate the parameters of the cameras. First, the internal parameters of the cameras are calibrated separately using Bouguet's calibration toolbox [10]. Then each neighbored camera pair are calibrated to obtain the external parameters.

Our color calibration module consists of an off-line calibration step and on-line calibration step. Color calibration is an important step in the system. The goals of color calibration is to eliminate the color flicker in the rendering result and more important, to improve the compression efficiency among views. Since the color will change unpredictably with the external illumination, temperature and distance, thus white balance function for all the cameras must be turned off and online calibration is required for views which must inter predicted. On-line calibration between cameras connecting to different producers is impossible since PCI bus is too limited for all the 64 raw video to stream to a central calibration machine, thus our on-line calibration is restricted to the four cameras on the same PCs. Because there is a great overlapped region among the four views, we only modify the brightness of the views to make the average brightness of the overlap region equaled during on-line calibration. We go on off-line calibration described in [6] for all the cameras before the system is start up.

3.2. Real time light field compression

A dynamic light field (DLF) is a 5 dimension signal (2 dimensions in each image, 2 dimensions across planar cameras and 1 dimension in time coordinate) which shows great challenge to data storage let alone the requirement of IP network transmission.

As an effective and widely used algorithm, inter-frame prediction is used by most video coding approaches. The main disadvantage of this scheme for DLF transmission is that when switching to a particular region, decoding of the new streams may meets up with a P or B frame which force the decoder to wait for the next I frame. This seriously blocks the rendering operation. To guarantee just-in-time rendering, all of the encoded DLF must be streamed to any of the users. The work in [5] uses MPEG2 to compress DLF for storage purpose.

Lots of the light field compression researches [11, 12] focus on the fully exploitation of the coherence among the four spatial dimensions. If cameras are located dense enough, these methods can provide good compression ratios. However, in practical applications, it is often impossible to capture such a dense and ideal DLF, and because the defect in camera calibration, the coding efficiency is worse than temporal prediction [13].

The MPEG 3DAV group and the work of [14] are currently investigating compression approaches based on simultaneously temporal and spatial encoding. However, these schemes are often too complex to be used in practical transmission systems and are not convenient for view switching purpose. Our system uses a novel coding structure to deal with these difficulties.

Figure 3 illustrates our coding scheme. In the DLF, the images captured at the same instant constitute a static light field which is similar to a ”frame” in video coding. We define 2 types of light fields, and they are I field and P field. The traditional temporal prediction chains can be break and successive prediction correlation will be eliminated using the corresponding image in the former I field as the reference image for each image in P fields and the later I field. The compression efficiency of such prediction is still high since the camera array and the background of the scene are both static. For I fields, compression can be realized by coherence exploitation within individual I field and between I fields. The right of figure 3 shows the prediction chains in an I field where three of the four images which connected to the same producer PC are predicted from the other image. Inter correlation between I fields is exploited as shown in the left of figure 3.

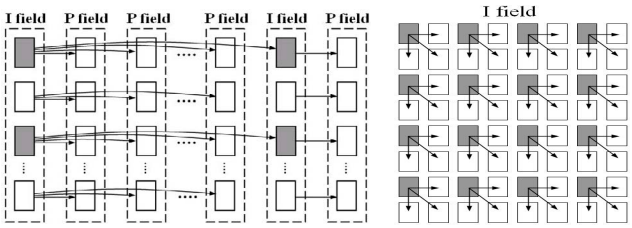

Fig.3. Prediction structure for the DLF compression: the left is temporal prediction structure between fields and the right is the spatial prediction structure inner I field

We implement this coding scheme through the modification of the Mpeg4 XVID codec. Note that such coding scheme does not involve multi-hypothesis prediction, and there is no information exchange between producer PCs. All these make our coding scheme suitable for real time DLF compression. Certainly, the total DLF can be stored in the producer PCs at this step. As for transmission, I field is imperative for every clients while images in P field can be selectively transmitted, thus our coding scheme can guarantee the multi-view simul-switching requirement.

3.3. DLF streaming and rendering

Figure 4 shows the block diagram for our light field transmission. Once there is client request for a streaming service, all the producer PCs send their 4 compressed streams to the streaming server and the stream server buffers the latest 2 second content in the memory. As users change their viewpoint and view direction, the client computes the view-streams required for rendering and sends the request through the feedback channel to the streaming server. Once the streaming server receives the message, it switches the P field streams and sends them through the data channel to the clients. Here, I field streams are independent to the stream selection and are all transmitted. The controlling delay user experienced may approximates to only the round trip time of message feedback. We use the dynamically light field rendering scheme described in [6] as our DLF rendering scheme.

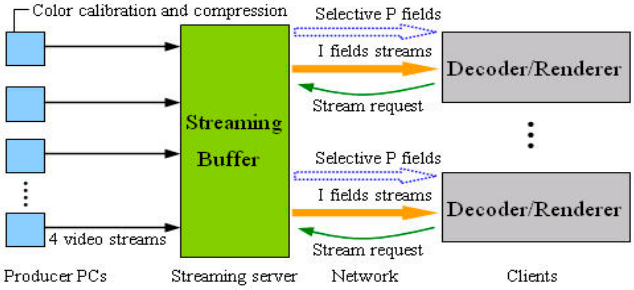

Fig.4. Block diagram for light field transmission

4. EXPERIMENTS AND DISCUSSION

Several tests have been carried out during our construction of the system. In the following, we will examine some system components including DLF capture, compression efficiency, network streaming and the loads of both server and client.

Figure 5 shows half of the views (4×8) in the first field captured by our light field cameras and figure 6 shows the 3 new views rendered on the client when the user focusing on different objects. There are some artifacts in the views which is the characteristic of light field system since the light field is not dense enough and the geometry calibration is not ideal in reality. However, the seamlessly roaming of the views gives users a 3D visual feeling.

Fig.5. Half of views in the first field of DLF sequence taken with our light field camera

Fig.6. Rendering focuses on different objects. The left focuses on the two men, the middle focuses on the door and the right focuses on the bear.

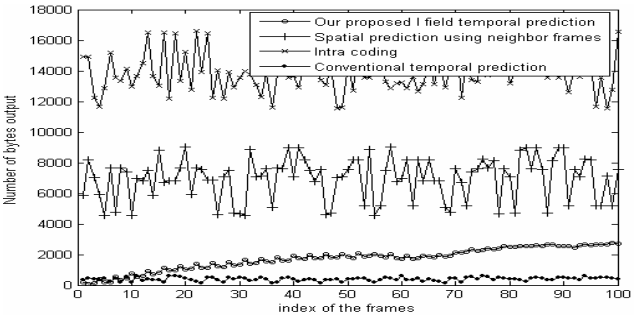

Fig.7. Comparison of the compression efficiencies using our DLF stream

Figure 7 compares the efficiencies of intra prediction, temporal prediction, spatial prediction and I field prediction. Here, all predictions are implemented with only one reference frame. Although the color calibration is satisfactory, the spatial prediction efficiency is still much lower than temporal prediction. For our proposed coding scheme, with the increasing interval between I field and the coding frame, the prediction efficiency becomes lower but still shows better performance than spatial prediction efficiency and approximates to conventional video coding. Figure 8 shows the simulation results of the required transmission bandwidth for each user versus the average number of views for rendering when the average image quality is in the range from 35.8dB to 36.1dB. We can see that when the number of views used for rendering is smaller than 16, our scheme outperforms the other two which are commonly used in the former works. Often, view points are set near the camera plane and only 4 views are required for rendering. The transmission bandwidth in this case is 2Mbps which is suitable for internet users.

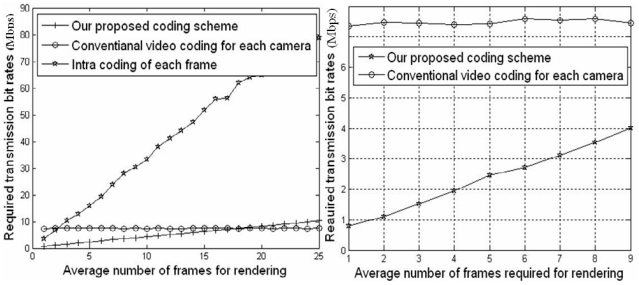

Fig.8. Average required transmission bandwidth changing with the number of views required for rendering. The left is the comparison of our coding scheme, the intra coding scheme and the conventional video coding scheme. To see the detail bit rate, the right is the rooming in of the left.

Using this system, we provide streaming service in the university's network. 9 views are used for rendering and the user felt delay is only 0.1~0.9 seconds. The producer PCs are powerful enough for 4 camera capturing, color calibration and data compression. Any client that has Pentium-IV D CPU 2.4G, 1GB RAM and a suitable GPU can real time decode and render these 9 camera streams at 30 fps. As for the streaming server which has a configuration of Pentium-IV D CPU 3.0G and 1GB RAM, figure 9 illustrates the emulated computational cost changing with the number of users. Also, we have examined transmission quality as the number of user increases. The output of our streaming server is Gigabits network and when user number is lesser than 16, lost rate will be lower than 0.5%.

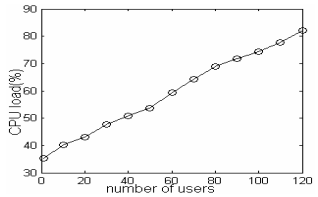

Fig.9. Computation cost vs. number of users

5. CONCLUSIONS

In this paper, we develop a real time interactive DLF streaming system for internet users. Real time data compression, free view point interactive streaming, arbitrary user support, 2D display and 3D display compatibility, minimized per user transmission bandwidth are all guaranteed in this system. We believe that with the development of camera and 3D monitor techniques, and as the computer processing power becomes stronger and networks bandwidth becomes broader, the vision experience of free view point control and 3D display over IP network will not be a dream.

6. REFERENCES

[1] M. Levoy and P. Hanrahan, "Light field rendering". *Computer Graphics (SIGGRAPH '96Proceedings)*, Aug.1996, pp.31–42.

[2] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph", *Computer Graphics (Proceedings SIGGRAPH-96)*, Aug.1996.

[3] Sloan, P., Cohen, M., Gortler, S. "Time-Critical Lumigraph Rendering." *1997 Symposium on Interactive 3D Graphics*, pp.17-24.

[4] MPEG Document, "Report on 3dav exploration", *ISO/IEC JTC1/SC29/WG11*, N5878, July 2003.

[5] Wilburn, B., Smulski, M., Lee, H.K., and Horowitz, M. 2002. "The Light Field Video Camera", *Media Processors 2002*, vol. 4674 of SPIE.

[6] Yang, J., Everett, M., Buehler, C., McMillan, L. "A Real-Time Distributed Light Field Camera", *Thireteen Eurographic Workshop on Rendering 2002*, pp. 77-85.

[7] S.C.Chan, K.T.Ng, Z.F.Gan, K.L.Chan, and H.Y.Shum, "The plenoptic video." IEEE Trans on CSVT, vol. 15, no. 12, pp. 1650-1659, Dec 2005.

[8] W.Matusik, H.Pfister, "3D TV: A Scalable System for Real-Time Acquisition, Transmission, and Autostereoscopic Display of Dynamic Scenes", *ACM Trans on Graphics (TOG) SIGGRAPH*, August 2004.

[9] J. Luo, H. Cai, J. Li, "A Real-Time Interactive Multi-view Video System", *Proc. of the 13th ACM International Conference on Multimedia(MM 2005)*, Singapore, Nov 2005, pp.161-190.

[10] http://www.vision.caltech.edu/bouguetj/calib_doc/

[11] 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

[12] B.Girod, C.-L. Chang, P.Ramanathan, and X.Zhu, "Light field compression using disparity-compensated lifting," in *proceeding of the IEEE International Conference ICASSP-2003*, April 2003.

[13] A. Smolic′, D. McCutchen, "3DAV Exploration of Video-Based Rendering Technology in MPEG", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 14, No. 3, pp 348-356, March 2004.

[14] MPEG document, "Survey of Algorithms used for Multi-view Video Coding (MVC)", *ISO/IEC JTC1/SC29/WG11*, HK, China, Jan. 2005