

A User-driven Interactive 3D Video Streaming Transmission System with Low Network Bandwidth Requirements

Zhenli Zhou, Li Zhuo, Jing Zhang, Xiaoguang Li

Signal and Information Processing Laboratory, Beijing University of Technology, Beijing, China

zhouzhenli@emails.bjut.edu.cn, zhuoli@bjut.edu.cn, zhj@bjut.edu.cn

Abstract— With the rapid development of electronic and computing technology, a growing number of new applications in the area of 3D video have come into sight, including the popular interactive 3D service. In this paper, the system architecture of a user-driven interactive 3D video streaming transmission is presented, which is simple to be realized with legacy displays. In particular, a novel idea is proposed, in which the viewer's position is determined using face detection and tracking technology, and then the location information is employed as feedback to the video server-side in real time, and at last the server adaptively selects the best matched parts of video stream data with the user's viewing angle. Under the existing network bandwidth conditions, the proposed system enables the viewers to select their viewing directions at their willing and enjoy a high quality of 3D experience anywhere they like.

Keywords— User-driven, 3D video, Interactive technology, Face detection, Stream switching.

I. INTRODUCTION

Three-dimensional interactive video technology is a new application in the field of video processing, which enables the viewers to freely switch the viewing angles to gain pleasing 3D viewing experience[1-2]. In this application, multiple video sequences of a scene should be captured at the same time, by cameras from different locations. To provide interactivity, the knowledge of the viewer's location is needed for such a system to select the correct two videos for the display equipment to create an illusion of 3D.

A typical approach to enable this kind of interactive 3D switching is to encode all the videos in one server site, and transmit all the streams to the remote ends. After the process of decoding at receiver side, the users can arbitrarily select two of the reconstructed videos they like better to enjoy the 3D perceptual effect. However, taken the cost of the systems into account, it is not applicable, under the existing hardware and software condition, to transport all the captured videos from different view points to the clients, as large network bandwidth is essential for transmission. Thus, this approach is difficult to be realized in the existing network environment.

As the key technology in the offered interactive 3D is an imitation of the human eye imaging system, where two 2D

views are shown separately to the viewer's left and right eye, it is common that only two adjacent captured videos that match the viewer's location make sense, leaving all the other video streams useless. Therefore, an alternative solution is proposed which sends only the desired streams to the user which requires the server to get the viewer's current position from remote end in an instant.

A client-oriented interactive 3D video system is proposed in [3-4], which adopted a scalable video coding structure to encode the video signals and divided the stream into base layer and enhancement layer. The base layer contains the low-quality stream data from all the views, and MVC (Multi-view Video Coding) HBP prediction structure with large quantitative parameters is utilized to achieve high compression performance. In addition, to realize random switching between various view points, the enhancement layers of each view is encoded independently and then combined with base layer, and thus the selected view's enhancement layer can improve the rendering quality of the view. At the same time, in order to facilitate interaction, the user's head position is tracked and predicted in the next time interval. Then both the predictive and practical position will be feedback to the video streaming server so as to adaptively select the best matched streams to the user's current viewing angle. The system will consume a large amount of network resources, and is sensitive to the prediction error of head information, which means that if there is prediction error or transmission errors in the enhancement layer information, only the MVC base layer could be decoded at the client. As a result, the viewer will get nothing but a low quality 3D video before the arrival of the correct enhancement layer data.

Another related system in this area is discussed by a group at Microsoft [5-7]. The key idea of the framework is to have the remote user send its current viewpoint and motion information to the capturing site, enabling it to provide the accurate virtual view and generate a set of weight maps to control an adaptive compression scheme, which will maximize coding efficiency and economize network bandwidth. The system is mainly for immersive teleconferencing applications, which can provide the viewers with a 2D image effect at any viewpoint. To synthesis the

desired virtual view in client-side, additional videos and depth maps should be encoded and transmitted to the remote end.

In this paper, a new user-driven interactive 3D video streaming transmission system is proposed, in which each view data of the multi-view video is separately encoded by the H.264 encoder and stored at the server side in advance. The face detection and tracking algorithm is adopted to determine the viewer's position at the remote side, and the location information will be feedback to the server side in real time. The two best matched video stream data will be adaptively selected at the server and then sent to the appointed client site. When transmitting to the remote end, only the two desired streams data are sent across the network for decoding and displaying. Therefore, under current network bandwidth conditions, the system is capable to provide the users high quality 3D videos in their pleasing view directions and achieve a better 3D perceptual experience.

II. SYSTEM DESCRIPTION

A schematic overview of the proposed system can be seen in Fig.1. The system can be divided into two modules: server and client. The multi-view video of N views is encoded (each view data encoded independently using H.264 encoder) firstly and stored on the server disk. At the client side, human face is detected and tracked in real time, and the results of face location will be transmitted to the server side. With that, the server determines the two best matched views data stream to be transmitted, and finds out the corresponding access point to fetch the stream data parts for transmission.

As is illustrated in Fig.1, the system includes several important parts, such as multi-view video coding, stream switching, face detection and tracking et al. Detailed descriptions will be presented as follows.

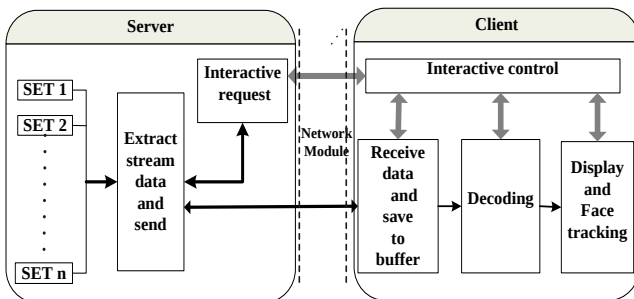


Figure 1. Overview of the user-driven 3D video streaming transmission system.

A. Multi-view video codec

Efficient multi-view video encoder/decoder is one of the key technologies in an interactive 3D video system. For multi-view video, frames in a single point of view possess time and space correlation. Besides, the corresponding frames between different viewpoints are also related with each other. In order to improve the compression efficiency, it is necessary to make full use of time, space and inter-view redundancies [8]. Currently, the multi-view video coding method has become one of the hot spots in video coding technology and various coding schemes have been proposed. Multi-view video coding

(MVC) could take full advantage of the space redundancy to improve the entire coding performance and the main idea is somewhat similar to the motion compensation in the traditional video coding frameworks that will inevitably produce dependencies between the motion compensation video frames. Hence, although the MVC prediction structure renders a high efficient compression, the coding correlations in viewing points brought by it makes the multi-view video difficult to provide random access. Therefore, MVC can achieve a high coding efficiency at the cost of high computational complexity and large memory requirements. Generally, it will be hard to meet the real-time decoding requirements, unless faster algorithms are brought out in the future.

Taking computational complexity, network bandwidth and other factors into account, the idea to use two view videos to provide the users the stereo perceptual experience directly is proposed in this paper. On one hand, it can easily render a high-quality 3D effect to the clients. On the other hand, the system could be implemented under the current network conditions since our goal is to significantly reduce the network bandwidth requirements. However, it is not suitable for the MVC coding method because of its complex view-dependency structure [9] (as depicted in Fig.2), which will make it very hard to randomly access the compressed bitstreams and decode in real time.

Therefore, in this paper, the H.264/AVC codec is employed to encode the multi-view video separately rather than the MVC encoder. Even though the overall video compression efficiency is somewhat lower compared to MVC and slightly larger memory is required to store the encoded stream at the server side, considerable bandwidth saving can be obtained. At present, there exist a variety of H.264 fast algorithms, which enables the proposed system easy to be implemented under current technique conditions. Moreover, the proposed separately encoding scheme supports random access to the stream.

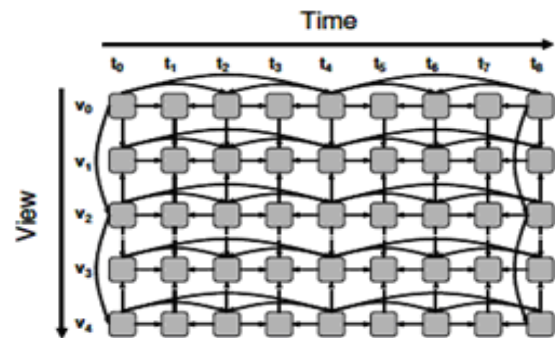


Figure 2. Sample prediction structure for multi-view video coding that exploits both temporal and spatial redundancy.

B. Stream switching and synchronization scheme

The major feature in this proposed interactive 3D video streaming system is that as long as the location of the user's face shifted at the client side, the server will be able to sense the change in time, and then adaptively select the best

matched video streaming data to be transmitted. Obviously, attention should be paid to the criteria of how to select the best matched streams according to a given position. The relationship of viewer's location and its corresponding videos will be briefly described as follows. As shown in Fig.3, a set of cameras are used to capture the scene at the same time. In order to simulate human binocular vision, the two adjacent cameras are spaced with fixed interval. Therefore, if the viewer eye position located within the interval region of two cameras, the two best matched view videos should be the ones captured by the most nearby cameras. As an example in Fig.3, if at some time the viewer's eye position P locates among the interval area of camera2 and camera3 which is marked with Area23, then the best matching left views should be V_2 and V_3 , which means that the videos from Cam2 and Cam3 should be transmitted the client sides.

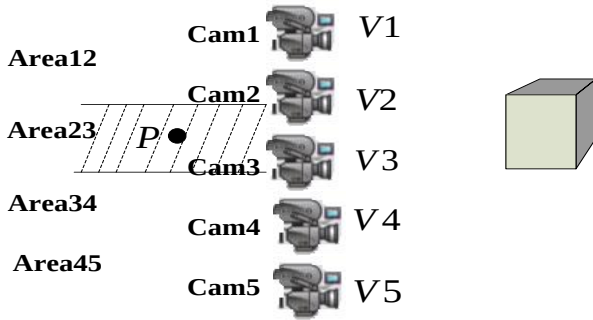


Figure 3. Relationship between the viewer's location and its corresponding matched videos.

Another main care is focused on the stream switching mechanism at the server site. Here, the mechanism should solve the synchronization problem, including two modes, time synchronization and view synchronization, details of which are shown in Fig.4.

Suppose that at a certain time, the sever is going to fetch the next frame data stream of view 0 and 1, but the interactive request of view switching to view 1 and 2 is answered. At this point, in order to ensure smooth video decoding and playing at the client side, the sever will firstly switch to the view 1 and 2, and then locate the corresponding time points in the view 1 and 2. Because H.264 decoder can not decode p frame data stream directly, the server need to locate the next I frame which is $I_1(i)$ and $I_2(i)$ in Figure 4 in the data stream, and send the following data streams to the clients. As for the viewpoint switching, it means to ensure smooth switching among the views. As shown in Fig.4, viewpoint switching requires views to transmit smoothly from the frames $I_0(l)$ and $I_1(l)(i < l < j)$ to $I_1(i)$ and $I_2(i)$.

The multi-view videos are encoded independently by H.264 in our proposed system, and thus there is not any correlation among the data streams from different views, which makes it easy to switch smoothly among the views. The H.264 bitstream's random access mechanism provides a good guarantee for the time-point switching within one view. The time delay problem during switching process can be solved by

selecting an appropriate period of I frame during the encoding process.

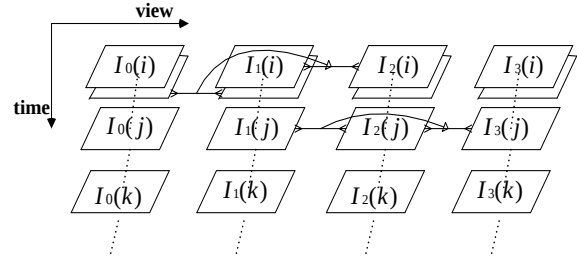


Figure 4. Stream switching in multi-view video streams.

C. Face detection and tracking

In order to get the user's viewpoint, the system proposed in this paper utilizes a video camera to capture the viewer's face movement in real time. The camera is fixed at the client side in advance. Relationship between the location of the user's face and its corresponding viewpoint is depicted in Fig.5. As shown in Fig.5, the 2D monitoring area of the camera is fixed. If a user's face is detected at the position $P(X,Y)$, then the best matched viewpoints can be determined between view V_i and V_{i+1} . According to this idea, the user's face is detected and tracked in real time at the client side, and the head position information will be sent to the server side. With the help of the face location information, the server can adaptively transmit the corresponding data streams from two views to the clients.

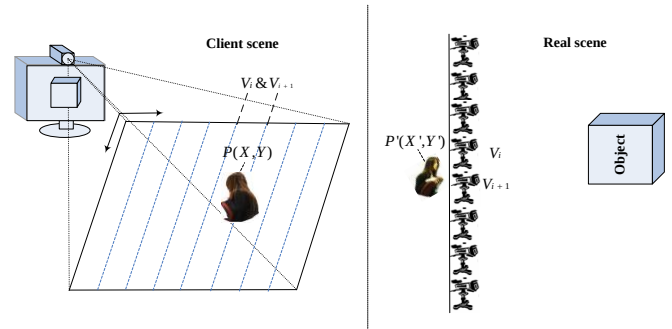


Figure 5. Relationship between the user's face location and its equivalent practical viewpoint.

III. SYSTEM IMPLEMENTATION

A. User application interface

A screenshot of the running system is illustrated in Fig.6, where the right-bottom part is the face detection and tracking window. In the system-testing experiment in the LAN environment, the switching delay can hardly be felt when the location of the face shift from one area to another. Users can specify the IP address and port information in the interface to connect to the server and take operations such as play, stop, turn on and off the face tracking window, exit and other manipulations during the watching process.

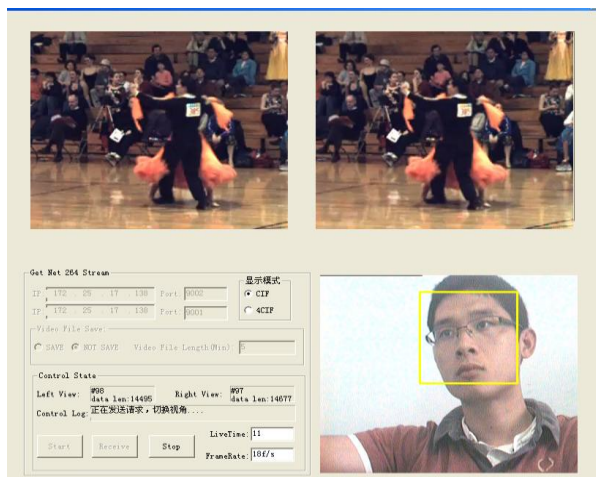


Figure 6. Client-side application interface of the user-driven 3D video streaming system

B. Face detection and tracking for real time applications

Human face is a nature structure with highly complicated variations in detail, which brings great challenges to the detection and tracking performance for real time applications. These kinds of variations lie in pose, facial expression, partial occlusion, lighting condition, rotation, etc. To solve these problems, a multi-pose human face method is employed. Combined with the multi-feature tracking algorithm, this face detection and tracking method enables the system to effectively determine the face area and quickly get the location of the human face. Experimental results show that the algorithm responds to a change of the face position very fast, which can meet the requirement of the practical applications. Fig.7 shows the results of the face detection and tracking in various conditions.

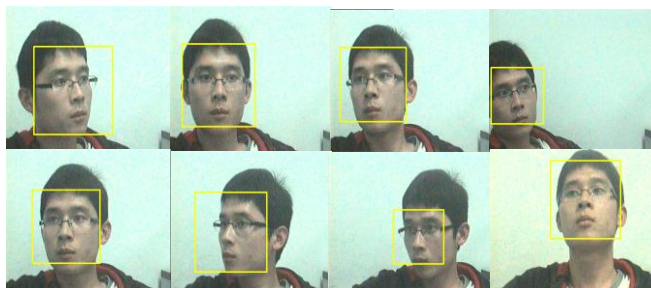


Figure 7. Results of the face detection and tracking system in various condition.

C. Implementation of the system control mechanism

The proposed user-driven interactive 3D video streaming system is required to deal with packet receiving and sending, fast decoding and synchronized displaying in real time. To ensure normal operation, a certain amount of data buffers are allocated at both the server and client sides. After fetching one frame data stream, the server will put some supplement information, such as views ID, frame type to the stream data, together with the data stream into a packet to be sent.

Similarly, at the client side, after the packets received, the left or right view data will be separated first, and then be put into the corresponding buffer to be decoded and displayed. Multi-thread approach is adopted at the client side to deal with data receiving and sending, decoding, displaying and face detection and tracking.

IV. CONCLUSIONS

In this paper, a user-driven interactive 3D video streaming transmission system is proposed. Thanks to the face detection and tracking at the client side, the system transmits only the best matching view data to the users, greatly reducing the consumption of the network bandwidth resources. In order to provide a high-quality stereo perceptual effect and smooth stream switching capability, the multi-view videos are encoded independently by H.264 in the system. Experimental results have shown that the system enables the user to choose his/her viewing angle freely, and can always provide the viewers with the suitable viewing videos.

ACKNOWLEDGMENT

The work in this paper is supported by the National Natural Science Foundation of China (No.61003289, No.61100212), the Natural Science Foundation of Beijing (No. 4102008), Supported by Program for New Century Excellent Talents in University, the Excellent Science Program for the Returned Overseas Chinese Scholars of Ministry of Human Resources and Social Security of China, Scientific Research Foundation for the Returned Overseas Chinese Scholars of MOE.

REFERENCES

- [1] ISO/IEC JTC1/SC29/WG11 N555822003 Report on status of 3DAV exploration [S].
- [2] ISO/IEC JTC1/SC29/WG11 N555822003 Applications and requirements for 3DAV[S].
- [3] K. Hopf, P. Chojeci, F. Neumann, and D. Przewozny, "Novel autostereoscopic single-user displays with user interaction," Proc. SPIE, vol.6392, p. 639207, 2006.
- [4] E. Kurutepe, M. R. Civanlar, A.M. Tekalp. "Client-driven selective streaming of multi-view video for interactive 3DTV," IEEE Transactions on Circuit s and Systems for Video Technology, 2007, 17 (11) : 1558~1565
- [5] J.-G. Lou and H. Cai and J. Li, "A Real-time Interactive Multi-view Video System", ACM International Conference on Multimedia, November 6-11, 2005, Singapore: I-161~I-170.
- [6] C. Zhang and Z. Yin and D. Florencio, "Improving Depth Perception with Motion Parallax and Its Application in Teleconferencing," IEEE International Workshop on Multimedia Signal Processing, Oct 5-7, 2009.
- [7] D. Florencio and C. Zhang, "Multiview video compression and streaming based on predicted viewer position," in ICASSP, 2009
- [8] P. Vassiliadis, Z. Vagena, S. Sk Iadopoulos, et al. "Towards the modeling design control and execution of ETL processes," Information Systems, 2001, 26 (8): 537~561 (2002) The IEEE website. [Online]. Available: <http://www.ieee.org/>.
- [9] A. Vetro, Yea. Sehoon, M. Zwicker, W. Matusik, H. Pfister."Overview of Multi-view Video Coding and Anti-Aliasing for 3D Displays." Image Processing, 2007. ICIP 2007. IEEE International Conference , Sept. 16 2007-Oct: I-17~I-20.