Internet-Wide Multi-Party Tele-Immersion Framework for Remote 3D Collaboration

Zhong Zhou, Xiuwen Chen, Lin Zhang, Xuefeng Chang

State Key Lab. of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China

School of Computer Science and Technology, Beihang University, Beijing 100191, China

ABSTRACT

Tele-presence and tele-immersion systems are growing to be more realistic these years. However, few show practical experiences in Internet wide tele-immersion, especially in acceptable end-to-end delay. We present a framework for multi-party tele-immersion applications that allow remote collaborative 3D interactions. This framework performs real-time stereo modeling based on the bumblebee cameras. It provides an architecture that supports several tele-immersion nodes to communicate in groups. This paper addresses how to reconstruct, compress and establish the system. The system has been set up in China NGI. Preliminary experiment results show that it's possible to have the delay under 100ms with 320*240 depth map and 640*480 texture images. The framework is also expected to be further extended for interaction with full-body reconstruction.

KEYWORDS: Tele-immersion, real-time modeling, compression.

INDEX TERMS: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism —Virtual reality; I.4.2 [Image Processing and Computer Vision]: Compression (Coding); I.4.5 [Image Processing and Computer Vision]: Reconstruction

1 INTRODUCTION

The technologies of 3D display have attracted a lot of enthusiasms in recent years [1]. Among all the 3D display applications, many people proposed a lot of imaginations on having persons feeling each other in a shared virtual environment in real-time. The two blockbusters, MATRIX and AVATAR, are both derived from idea of this topic. Since 1990s, some commercial products have been put forth in the field of tele-presence or tele-immersion, e.g. the TelePresence system from Cisco, the Halo system from HP Lab etc. These products are aiming at the multi-party video conference. On the mean while, researches on more realistic 3D collaboration are ongoing.

Tele-immersion is referred to as adding some kind of interactions to the simply tele-presence so that one can "touch" the shared virtual environment besides merely the 3D reproduction of the human body. Modern tele-immersion technology is mainly based on multi-view images captured from cameras in different directions to an object space. It conducts real-time 3D reconstruction from the images and then delivers the model with textures to the rendering or other distributed sites via network. All the work need to be done in one frame to obtain real-time 3D perception. Obviously, this technology is a multi-disciplinary topic related to computer vision, graphics, network and real-time data processing etc. Tele-immersion reconstructs the objects in the real world in real-time, as starts a new kind of Virtual Reality interaction that mixes the real objects with the virtual environment.

Some work has been done in the tele-immersion field. Though, few show practical experiences in Internet wide tele-immersion to now, especially those of acceptable end-to-end delay. There are two difficulties inside tele-immersion. They all lie on the processing time that affects the end-to-end interaction delay. One is the efficiency of 3D content generation including reconstructing the mesh and extracting/compressing the high resolution texture. The other is the transportation efficiency since the size of the models refreshing several times each second will be very big. A real-time stereo vision based tele-immersion framework is presented in this paper. There are two main contributions: first, the framework designs the communications among the users and the TI (tele-immersion) server that provides a platform for remote 3D collaboration in groups; second, with our techniques, we developed the system and pursued tele-immersion experiments over the Internet, which may be useful for related reference. Some of our techniques in the system are also introduced including the stereo vision based 3D matching and the real-time compression on the dynamic depth map with high resolution texture images.

2 RELATED WORK

There exist some well-known multi-camera based tele-immersion systems, such as GrImage and successive Vgate by INRIA [2], TI and TEEVE systems by U.C.Berkeley, U.C.Davis and UIUC [3], SVTE by the Heinrich-Hertz institute [4]. These systems all recover the 3D information from multi-view cameras, make real-time 3D reconstruction, and then apply the results to the mixed reality environment. The real-time reconstruction plays the key role in this process since the algorithms need heavy computation to generate the 3D mesh and texture in a short time no more than 0.5 second. Although the newly coming time-of-flight (TOF) cameras are available in the market from the ZCam (now belongs to Microsoft), mesa, PMD or Canesta companies, their resolutions are always very low to below 200*200 as is not sufficient for full-body interaction. With the resolution improvement of TOF cameras, it will be sure to change the traditional reconstruction algorithms. The real-time reconstruction algorithms are mainly of two kinds. One is the stereo vision based reconstruction which calculates the depth of pixels based on the 3D matching of the pixels in two images. The other is based on the structure from silhouette (SFS) in which visual hull algorithm is now a hot topic. The stereo vision based reconstruction adapts to various of object shapes, but it has to compromise in the matching precision and overhead with different baseline. The Berkeley TI, SVTE are based on the stereo vision based reconstruction. The visual hull algorithms can provide good quality full-body out-looking in some viewpoints by calculating the intersection of vision cones. It

Email Address: zz@vrlab.buaa.edu.cn

IEEE International Symposium on Virtual Reality Innovation 2011
19-20 March, Singapore
978-1-4577-0054-5/11/$26.00 ©2011 IEEE

has higher computation complexity and difficulties in concave surfaces. GrImage and Vgate are based on EPVH [5] of this kind. The visual hull algorithm is more costly because of the requirements on the size of capture space and number of cameras.

To now, little work has been done in the real remote sites except the Berkeley TI which shows high traffic and big delay in the experiments [5]. We established a tele-immersion system in 10 university sites of China Next-Generation Internet (NGI) to make experiments. The 3D reconstruction includes both the stereo-vision based and visual hull algorithms, but for now only the stereo-vision based reconstruction is integrated and tested. The visual hull reconstruction has a lot of traffic in sending multiple texture images that brings a big problem to our real-time remote communication. All the 10 sites are equipped with 3-camera bumblebee camera as Figure 2. The design and preliminary experiment results are provided in this paper. The data may provide some useful reference for related work.

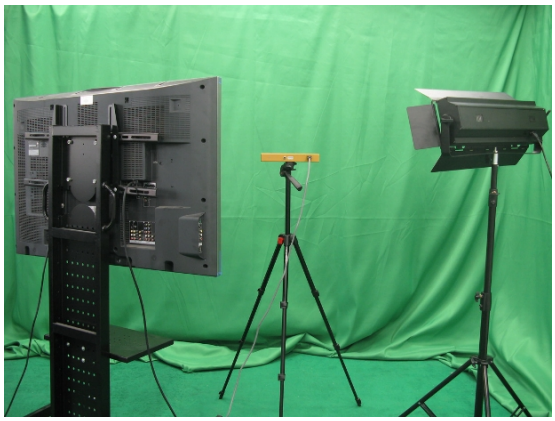

Figure 1. TI Client Settings

3 FRAMEWORK OVERVIEW

Our work of tele-immersion is tested in the China NGI to verify this new kind of remote interaction. Considering the system scalability, the TI system is divided into the server side and TI client as Figure 2. The server side includes the TI server, system portal and the database.

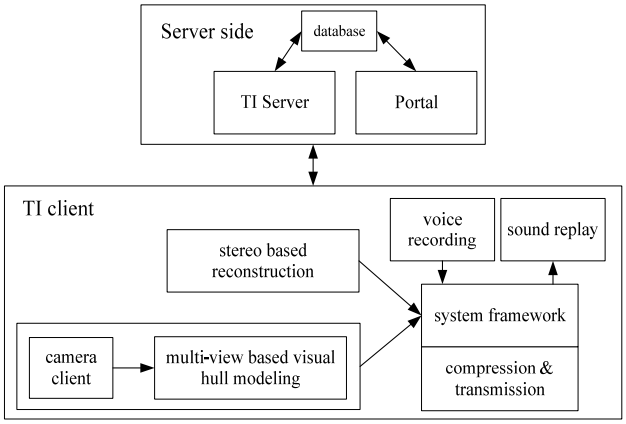

Figure 2. Tele-Immersion system architecture

The TI server provides a virtual Hall and room model to divide the users into groups. The system information such as user login/logout and their manipulations are logged in the TI server, and some of them are also written to the database. The portal provides user login and real-time status display through web browsers with the help of database. The TI client gets the 3D user data from reconstruction, stereo-based or multi-view based one, and performs data compression, transmission and rendering in the system framework. The framework deals with the user interaction, some physical simulation, and the 3D rendering. The TI client has some big problems inside such as the real-time 3D reconstruction and a scalable communication model.

Two types of real-time 3D reconstruction are implemented in the system including the stereo-based reconstruction and the silhouette-based visual hull reconstruction. For the stereo-based reconstruction, the remote client only needs one image per frame to perform texture mapping because actually the real-time model is based on a depth map. However, for the visual hull modeling, the images of all the directions are required for the full-body mesh. In our experiments we have 8-12 cameras covering a space, so images from all these cameras are all required in each frame. It's a heavy burden to the whole system. For now, only stereo-based reconstruction is implemented in our system. More efficient multi-view based texture compression algorithms are still under investigation since current schemes are not efficient in real-time situations.

In a typical multi-party tele-immersive environment like 3D video conference, each one sends and receives meshes and corresponding texture images to/from several other ones. It will incur heavy traffic and then affect the system scalability and performance. A communication model is designed as Figure 3 for the tele-immersion, in which the TI server coordinates the client communication. The TI server lists the user clients in groups each of which is associated with a virtual room. When a new client logins, the TI server will push the list including all the other clients in its group to it. The client will get connected with all the other partner clients via TCP or UDP. In this way, the clients will communicate their models, texture images and sound in real-time. We also tested this model with visual hull-based models. The TI clients will exchange the full-body meshes with the selected one or two texture images according to their viewpoints. Preliminary experiment results show that the performance slows down heavily due to the traffic and delay. Further work on the reconstruction and compression algorithms need to be conducted to meet the interaction requirements.

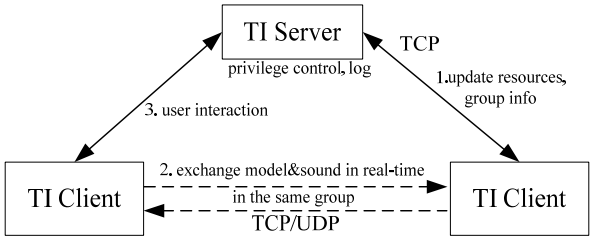

Figure 3. Communication model

Different with the mesh/texture data, the user interactions including the login/logout, manipulation are forwarded by the TI server. The TI client will launch one specific thread to download such initialization data as 3D scene resources and user group lists. The interactions are always required to be reliable delivered, so the connection between the TI client and the server is designed to be TCP. On the other hand, the 3D data including the meshes and texture images is updated in real-time. When some 3D frames are dropped, it's acceptable. Hence, the 3D data is possible to be transmitted in unreliable channel such as UDP.

4 REAL-TIME STEREO-BASED RECONSTRUCTION

Real-time reconstruction is the critical algorithm in tele-immersion systems. It's almost equally important to have the well reconstructed models and to be real-time to meet the user interaction requirements. For the real-time stereo-based reconstruction, it captures two images with disparity in each frame, makes stereo matching to calculate the depth map and then gets the mesh through triangulation. The stereo matching algorithm is the main part that affects the accuracy and efficiency of the reconstruction.

A real-time stereo matching algorithm RealTimeLAW is developed in the system that estimates scene depth information in real-time. The stereo algorithm consists of two steps.

First, the color similarities are used to compute a layered support weight for each pixel and to aggregate matching costs. Given a stereo image pair $\{I, \bar{I}\}$, where I and $\bar{I}$ are the reference and target images, the initial matching costs are computed using the absolute difference method:

$$C(p,d) = \sum_{c\in\{r,g,b\}} |\, I_c(p) - \bar{I}_c(\bar{p}) \,| \tag{1}$$

where I_C and $\bar{I}_c$ are the color intensities in the left and right images. Symbol p denotes the pixel (x, y) in the left image, and d is the disparity hypothesis. The corresponding pixel $\bar{p}$ in the right image is the one $(x+d, y)$. We build a square window W_p of predefined size centered at each pixel p. For each pixel $q \in W_p$, we define a weighting function $w(p,q)$ that computes a weight representing the likelihood to which pixel q lies on the same surface with pixel p. The adaptive support-weight is computed as:

$$w(p,q) = L(\exp(-\frac{\Delta c_{pq}}{\gamma_c}))\exp(-(\frac{\Delta c_{pq}}{\gamma_c} + \frac{\Delta g_{pq}}{\gamma_d})) \tag{2}$$

$$L(x) = (\frac{1}{N-1})\,floor(\frac{x}{1.0/N}) \tag{3}$$

where Δc_{pq} denotes the colour distance between p and q, Δg_{pq} is the Euclidian distance between p and q on the image plane, $floor(t)$ returns the nearest and lower integer value of t. Symbols γ_c, γ_d and N are constant parameters determined empirically. The aggregated cost is computed as a weighted sum of each pixel cost in support window. i.e.

$$C'(p,d) = \frac{\sum_{q\in N_p} w(p,q)\cdot w(\bar{p},\bar{q}))\cdot C(q,d)}{\sum_{q\in N_p} w(p,q)\cdot w(\bar{p},\bar{q}))} \tag{4}$$

where N_p is the set of all pixels in the support window.

Second, we use an amended scan-line optimization technique which combines winner-take-all (WTA) and dynamic programming (DP) to compute disparities. By combining WTA and scan-line DP, our approach can compute disparity of pixels in depth discontinuity areas. The formula is:

$$F(p,d(p)) = C'(p,d(p)) + B$$
$$B = \min_{d\in\{d(p)\pm1, d(p), d_{x-1}\}} \{F(p',d) + \gamma * abs(d(p) - d)\} \tag{5}$$

$$d_{x-1} = \arg\min_{d} C'(p',d) \tag{6}$$

where $p' = (x-1, y)$, $d(p)$ is the disparity function of pixel p. Symbol γ is a constant used to penalize depth discontinuities.

The experimental results are evaluated on the Middlebury benchmark data sets, showing that to now RealTimeLAW achieves the best reconstruction accuracy among existing real-time stereo algorithms. Table 1 shows quantitative results that are taken from the Middlebury online table.

Table 1. accuracy of real-time stereo matching algorithms

Algorithm	Avg Rank	Avg Error	Tuskuba nonocc all disc	Venus nonocc all disc	Teddy nonocc all disc	Cones nonocc all disc
RealTimeLAW	**34.6**	**6.56**	**1.40, 3.07, 5.86**	**0.73, 1.74, 3.86**	**6.81, 14.0, 15.4**	**3.99, 11.8, 10.1**
Adaptive Weight	33.9	6.67	1.38, 1.85, 6.90	0.71, 1.19, 6.13	7.88, 13.3, 18.6	3.97, 9.79, 8.26
RealTimeABW	41.2	7.90	1.26, 1.67, 6.83	0.33, 0.65, 3.56	10.7, 18.3, 23.3	4.81, 12.6, 10.7
RealTimeBP	45.2	7.69	1.49, 3.40, 7.87	0.77, 1.90, 9.00	7.78, 14.9, 17.3,	4.58, 12.4, 10.7
RealTimeVar	53.1	9.05	3.33, 5.48, 16.8	1.15, 2.35, 12.8	6.18, 13.1, 17.3	4.66, 11.7, 13.7
RTCensus	58.0	9.73	5.08, 6.25, 19.2	1.58, 2.42, 14.2	7.96, 13.8, 20.3	4.10, 9.54, 12.2
Real-Time GPU	59.1	9.82	2.05, 4.22, 10.6	1.92, 2.98, 20.3	7.23, 14.4, 17.6	6.41, 13.7, 16.5

5 REAL-TIME COMPRESSION

To obtain natural and prompt interaction, TI nodes need to get the 3D models in time. The 3D models include mesh triangles and textures. Textures are from the foreground pixels, so the size is very big. Like other image compression, real-time compression is essential for real-time delivery. Yang et al. proposed a lossy compression scheme for the depth data and a lossless one for the color data in tele-immersive environments [7], but those are frame independent compression, so the compression ratio is limited. In fact, as the tele-immersive environments only care about the foreground objects, the valid textures only exist in the foreground pixels. We proposed a Border-descript Inter-Frame compression (BIFC) scheme, which applies inter-frame motion estimation with the restriction of foreground object pixels. The scheme has been integrated to the Berkeley TI system [8].

Our reconstruction is based on stereo modeling, so each vertex of mesh is defined in a three dimension space (x, y, disparity). Its corresponding texture coordinates are set in two dimensions as a normalized form (x/width, y/height). Then, for each part of the stereo model, one frame is enough for the textures of the part. In this means the texture mapping can be conducted with GL_TRIANGLES rendering in arrays.

For the depth map, it's easier to pursue fast compression. For example, we use the run length encoding (RLE) algorithm for depth map compression. The compression ratio is usually about 7:1 in our experiments. Similar to all kinds of image compression, at the first stage the images are divided into macro blocks to increase the compression speed. The foreground mask will also be

derived after background subtraction during reconstruction. Three types of blocks are defined in our scheme, *A* for total foreground, N for background and *B* for Border i.e. not all the pixels inside foreground (Figure 4). Here the border block type *B* is extra defined compared with other bitmap based methods.

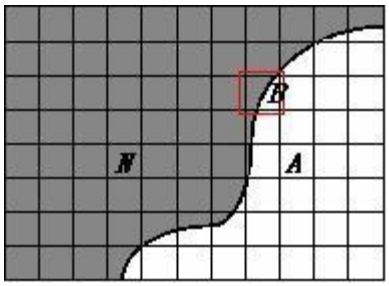

Figure 4. Foreground object blocks

According to the mask in pixel, all the blocks can have their types. Let $M[w,h]$ be the original mask of width w, height h. Let Macro Block has the size $d*d$. Define mask as $M[x,y]$, $(i-1)*d \le x < i*d$, $0 \le i < w$, $(j-1)*d \le y < j*d$, $0 \le j < h$. Let $M'[i,j]$ be the new mask in blocks. We have

If all $M[x,y] == 0$, *then* $M'[i,j] = N$.

If all $M[x,y] == 1$, *then* $M'[i,j] = A$.

If exists any $M[x,y] == 0$ *and any* $M[x,y] == 1$, *then* $M'[i,j] = B$.

The inter-frame compression uses reference frames for motion estimation. A group of frames begins with the reference frame and then the P frames. The reference frame is compressed with JPEG compression, called intra-frame compression. A modified JPEG compression is implemented which only compresses the border and inside blocks i.e. background blocks are ignored for no usage in textures. The intra-frame compression consists of color channel transform, DCT (discrete cosine transform), quantification, RLE and Huffman coding. The reference frame's encoding structure is a sequence of macro block type, mask encoding and the image encoding data.

The intra-frame compression with mask we adopted is not simply searching line by line. In stereo based modeling, the angle between the view directions of the two cameras is not big and the user won't act with large movement in most of the time. So the border set is used for a quick search because the border blocks have more distinct feature of having both foreground and background pixels. In this way some estimation with direction could be applied. We start from the first border block that is found during scanning and then search border blocks in its 9 neighbor blocks. Assume the valid border blocks is *k*. Let the size of search window be *u*u* where *u* is the edge length of the window. Then search in the search window of reference frame for the *k+1* border blocks with the smallest deviation. The block deviation is calculated by the regular equation:

$$dev = \sum_{i=0}^{d-1}\sum_{j=0}^{d-1}\left|I(i,j) - P(i,j)\right| \tag{7}$$

Then the moving vectors can be obtained from the paired blocks. The initial $\overline{MV}$ is set to be the average value of all the first several available block vectors.

$$\overline{MV} = \frac{1}{k+1}\sum_{i=0}^{k} MV_k \tag{8}$$

The entire foreground pixels may come from several objects with different moving pattern, $\overline{MV}$ only takes effect in the lines that begin with any border block in last border sets. After the lines, it has to be recalculated. The following block searching, DCT transform of block residuals are as those in JPEG. The P frame's encoding structure is a sequence of macro block type, mask encoding, image residual encoding and moving vector data.

6 EXPERIMENT EVALUATION

The tele-immersive system for research communication over CNGI has been developed. Some local and remote experiments have been carried out. The system GUI is as Figure 5 which provides tens of virtual rooms for each kind of research scenarios where the users can make tele-immersive communication. One of the 3D tele-immersion environment is as Figure 6 where three users making discussions on the protein molecule structure. We can see that the users can manipulate and interact with the virtual objects with their bodies immersed in the virtual environment. Figure 6 is a chess game in which people enter the chess environment and play with their bodies. Since the real-time 3D construction produces the dynamic model sequence, the collision detection algorithms based on the discrete time don't fit well and will incur the phenomena of penetration. The successive collision detection algorithms will among one of the future work besides the full-body 3D reconstruction. Here the preliminary experiment results are presented.

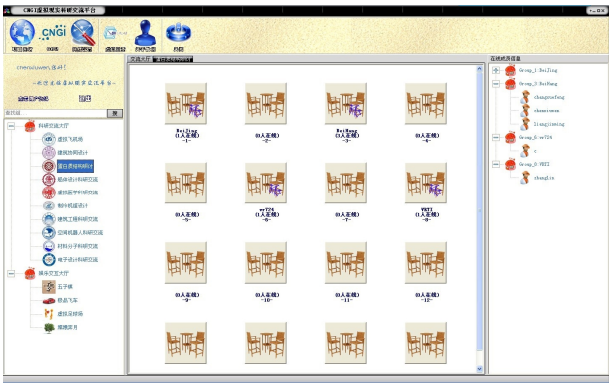

Figure 5. GUI of tele-immersive research system

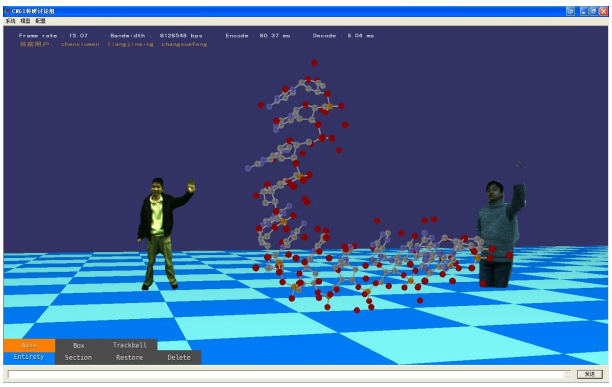

Figure 6. Collaborative tele-immersive discussion

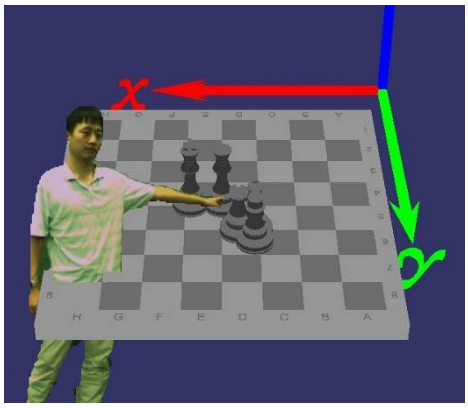

Figure 7. Interaction in a Chess Game

The TI client hosts are of 4-core Intel Xeon W3530 2.8GHz, 6GB DDR3 1333MHz ECC memory, 2*500GB 7200rpm SATA hard disk, and NVIDIA Geforce GTX 470 display card. They are all running with windows XP. The depth map is set to the resolution of 320*240 during the reconstruction, and the texture images have the resolution of 640*480.

Normally the time cost of 3D reconstruction & encoding ranks No.1 source of delay than the peer-to-peer network delivery. Figure 8 illustrates the time cost of this period. The time keeps less than 80ms which varies with the size of the foreground object. When the foreground object occupies a small portion, the time is smaller even to less than 50ms as Figure 8.

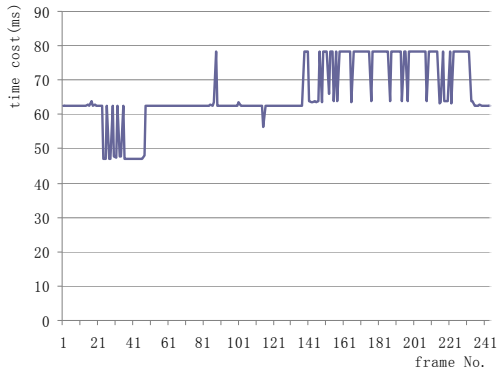

Figure 8. Time cost of 3D reconstruction & encoding

Figure 9 is the results of bandwidth occupation when 3 tele-immersive users in a virtual environment. The client bandwidth maintains about 8 Mb and the model update speed is about 14 times per second.

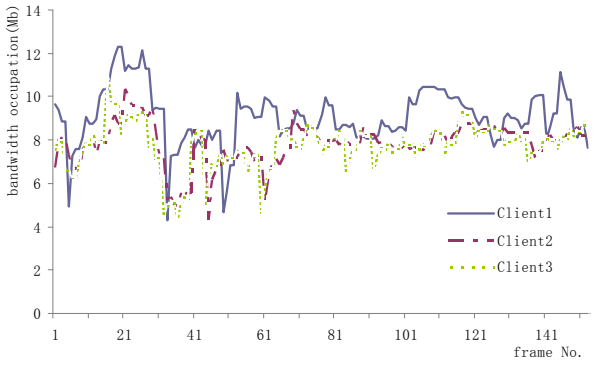

Figure 9. Bandwidth occupation (3 users)

Figure 10 is the results of the time cost of 3D model decoding and rendering. The time cost also varies with the size of foreground object. The time cost of this period is about 6-12 ms.

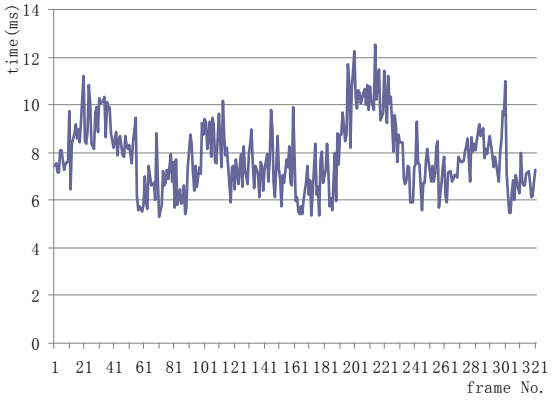

Figure 10. Time cost of decoding & rendering

The delay of user interaction on the virtual objects goes through the server, not in the peer-to-peer delivery. The preliminary experiment is conduct in two places of Beijing, Beihang Univ. and Beijing Univ. of Posts and Telecommunications. The result is as Figure 11 which shows that the delay is between 4-16 ms.

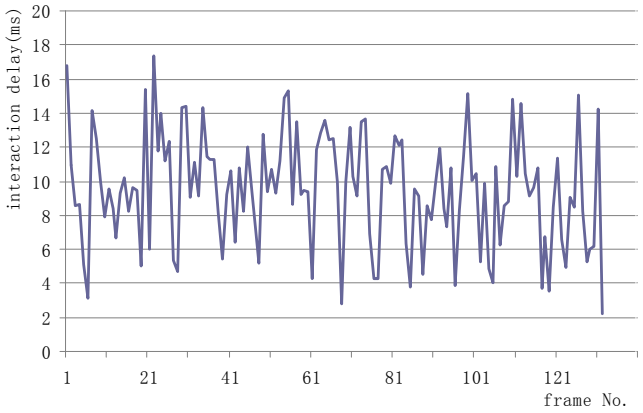

Figure 11. Remote interaction delay on 3D objects

Consequently, considering the time cost of 3D reconstruction & encoding and that of decoding & rendering, the total system delay is possible to be under 100ms when the delay among the remote places are small. The delay under 100ms is excellent for this kind of Internet collaboration. In fact, we have just completed the experiments among Tsinghua Univ.@Beijing, Beihang Univ.@Beijing and Huazhong Univ. of Science and Technology@Wuhan. However the node delays seem to be among 200-300ms as is far beyond our experiences 4 years ago in 2007. We guess the factors may come from the network authentication of Beihang Univ. or recent CNGI backbone testing etc. The problem is still under investigation. So the results aren't presented here. Further exploration will be carried out.

7 Conclusion and Future Work

Tele-presence and tele-immersion systems are growing recently. Our tele-immersive system over CNGI is presented in this paper including brief introductions to the system framework, real-time 3D reconstruction and compression algorithms. The framework is designed considering the communication model and scalability to support future multi-party communication in groups at the same time. The 3D reconstruction has to deal with the accuracy and

efficiency together. The idea of 3D compression algorithm is presented here which is also implemented in our previous work [6]. Preliminary experiment results show that it's possible to have the delay under 100ms with 320*240 depth map and 640*480 texture images. Future work will include the Internet-wide experiments, full-body 3D reconstruction, and mixed reality interaction etc.

ACKNOWLEDGEMENT

This work is supported by the Natural Science Foundation of China under Grant No.61073070, the National Grand Fundamental Research 973 Program of China under Grant No. 2009CB320805, the 2008 China Next Generation Internet Application Demonstration sub-Project under Grant No. CNGI2008-123, and Fundamental Research Funds for the Central Universities of China.

REFERENCES

[1] ZHAO Qinping. A survey on virtual reality. SCIENCE IN CHINA Series F: Information Sciences. 2009, 52(3):348-400

[2] Benjamin Petit, Jean-Denis Lesage, Edmond Boyer, Bruno Raffin. Virtualization gate. ACM SIGGRAPH 2009 Emerging Technologies, New Orleans, Louisiana, 2009

[3] Gregorij Kurillo , Ramanarayan Vasudevan , Edgar Lobaton , Ruzena Bajcsy, A Framework for Collaborative Real-Time 3D Teleimmersion in a Geographically Distributed Environment[C], Proceedings of the Tenth IEEE International Symposium on Multimedia (ISM), December 15-17, 2008:111-118

[4] Peter Eisert. Virtual video conferencing using 3d model-assited image-based rendering [C]. Proceedings of the 2nd IEE European conferenc e on Visual Media Production(CVMP),London, Nov 30-Dec 1,2005:185-193

[5] J-S. Franco, E. Boyer. Eficient Polyhedral modeling from silhouettes. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.31, no.3, pp. 414-427, 2009

[6] Zhenyu Yang, Wanmin Wu, Klara Nahrstedt, Enabling Multi-party 3D Tele-immersive Environments with ViewCast, ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP), vol. 6, 2010.

[7] Zhenyu Yang, Yi Cui, Zahid Anwar, etc. Real-Time 3D Video Compression for Tele-Immersive Environments, Proc. of SPIE/ACM Multimedia Computing and Networking (MMCN'06), San Jose, 2006.

[8] Ramanarayan Vasudevan, Zhong Zhou, Gregorij Kurillo, Edgar Lobaton, Ruzena Bajcsy, Klara Nahrstedt. Real-Time Stereo-Vision System for 3D Teleimmersive Collaboration. International Conference on Multimedia and Expo (ICME), July 19-23, 2010, Singapore:1208-1213