Tele-immersion Portal: Towards an Ultimate Synthesis of Computer Graphics and Computer Vision Systems

Amela Sadagic[1], Herman Towles[3], Loring Holden[2],
Kostas Daniilidis[4] and Bob Zeleznik[2]

1 amela@advanced.org, Advanced Network and Services, Armonk, NY
2 lsh | bcz@cs.brown.edu , Brown University, Providence, RI
3 herman@cs.unc.edu, University of North Carolina at Chapel Hill, NC
4 kostas@grip.cis.upenn.edu, University of Pennsylvania, Philadelphia, PA

Abstract

We describe a novel approach for unifying computer graphics and computer vision systems, and our initial results in building and using a prototype system. This approach has three significant characteristics: unification of the real and virtual worlds for both input and output, tele-collaboration between remote participants, and interaction between heterogeneous stand-alone 3-dimensional (3D) graphics applications [1]. The system is designed to run on the networks of the future, and it is capable of transmitting blends of dynamic computer graphics and computer vision data in real-time. In this text we concentrate on its visual part, in particular on synergy of computer graphics and computer vision systems as a new medium for collaboration and tele-presence. The preliminary steps of our research make us increasingly optimistic that in the future this technology will provide highly compelling, immersive environments for an increasing variety of tele-collaborative applications over high bandwidth networks

Keywords: computer graphics, computer vision, Virtual Reality, tele-immersion, Internet2[*]

Introduction

The ultimate goal for a large majority of researchers in the areas of Computer Graphics (CG) and Virtual Reality (VR) is a system that "looks real, acts real, sounds real, and feels real"[2]. The ideal system is a sensory-rich environment that provides a highly compelling experience that is the epitome of presence. The people involved in National Tele-Immersion Initiative (NTII, “Initiative” in further text)[**] have taken the preliminary

[*] Internet2 (http://www.internet2.edu/) is a consortium led by more than 180 members, most of which are universities and research institutions. The core of Internet2 network is its backbone called Abilene with OC48 (2.3 Gb/s) connections between its major nodes.

[**] National Tele-Immersion Initiative (http://www.advanced.org/teleimmersion.html) is a research consortium made of several research groups, led and sponsored by non-profit company Advanced Network

steps towards the same goal. In 1997 we started creating an application that will challenge major Internet2 network parameters (bandwidth, delay and jitter). The goal was to work on a system that would support remote human collaboration and extend our knowledge in the arena of intuitive user interfaces and human factors with the intent to build a tele-collaborative, immersive VR system.

In addition to this basic goal we established further requirements: the system should support an environment that looks like a natural and functional extension of the space that people use in their everyday work, and therefore it should not be placed in a special, dedicated space isolated from people's every day's activities. The entire system should be as intuitive and inclusive as possible, and it should require as little preparation for a collaborative session as making a phone call. Having all of this in mind we decided to situate our initial prototype of the tele-immersive system in the corner of an office where, in general, any (potentially irregular) surface in that space can be used as a spatially immersive display [3].

There are valuable lessons to learn from the systems that support human tele-collaboration and communication, and the most remarkable ones are Virtual Reality (VR) systems and videoconferencing systems. The studies of VR systems [4][5] suggest several factors that are important for the sense of presence (feeling of being in an environment that is different from the physical environment) and copresense (sense of togetherness and sharing the same environment with other users). They are 3D life-size, stereo, photorealistic-looking data, high refresh rate for all kinds of sensory data used in the system, low latency, wide field of view, ability to interact with system and with other users using intuitive graphics interfaces, existence of personal (individual), egocentric or exocentric viewpoint for each user, ability of the user to move about freely, the existence of the system that tracks those movements and updates images accordingly with a minimal time lag, to name just a few.

On the other hand the videoconferencing systems deliver photographic images of the real environment through its visual communication channel, and deliver a faithful replica of the real world. However, the features missing from such systems offer important lessons too: true eye contact, gaze awareness, and ability to view remote environment from any angle and direction. In addition to these feature, numerous studies of both VR and videoconferencing systems emphasize the significance of having very good speech communication channel and audio in general, preferably delivering 3D (spatialized) audio to each user (participant) in tele-collaborative session [6].

Representation of Humans: Avatars versus 3D Acquisition

The inevitable issue in any tele-collaborative VR system is a form in which humans and their acting will be represented in the system. In general, there are two ways of representing humans in such systems: using avatars (3D computer graphics models as

---

& Services, Armonk NY. Current university partners on the projects are Brown University, University of North Carolina at Chapel Hill and University of Pennsylvania

approximations of humans' embodiments) or using outputs of different 3D acquisition systems (real time 3D "scans" of humans).

In order to provide real-time simulation of humans, typical VR system using avatars requires remarkable amount of computational power for the physically based modeling and real-time simulation of the range of human movements and gestures, human skin, hair, textures, clothes and any other detail that might be necessary in particular context. (It would be preferable to have full simulation of humans but some applications might not impose such high requirements. These algorithms are extremely hard and they never cover the entire universe of possible complex gestures and movements so typical for humans.) This approach requires building elaborate 3D models of humans in advance, devising a set of non-trivial algorithms that would perform required simulations, tracking human activities and, at the end, simulating those and presenting them in a visual form to the users, all requested to work in real time. Real-time physically based simulations are very difficult tasks because of their huge computational demands. Because of the limitations present in computer graphics technologies the quality of simulated worlds is still far from real, which is particularly noticeable in simulations of humans. It is extremely hard to represent fine and small details of human appearance and communication clues necessary to achieve a high sense of presence, such as subtle facial gestures, as well as the representations of skin texture and elaborate wrinkles [7], [8]. As a direct consequence people can still easily identify synthetic models and behavior. In a collaborative environment our trust and engagement in communication will depend on whether we believe something represents a real person or not. This, in turn, may have considerable effect on our task performance. In addition to all of this, the process of producing good and inevitably complex 3D model of the real world takes a lot of time and skill, something that is not available everywhere and not with the promptness that is usually needed.

Opposite, 3D acquisition systems invests all computational power not into understanding how the real world works and simulating it, but rather scanning and obtaining 3D visual representation of the same physical world, and reconstructing the closest 3D replica of that world as is. The same system and algorithm is applied on any physical object, and therefore the entire system has more generic approach. Here, again, all stages of the algorithm have to be performed in real time in order to extract 3D geometry from the objects that are moving and changing dynamically over the time. Therefore, 3D models of the world in such systems are acquired on the fly eliminating the need to develop any 3D model a particular application will need in advance. These algorithms, as much as algorithms for simulation of humans, are non-trivial. The real-time requirement always poses additional burden to any system and this makes the task even more difficult. The acquisition systems usually produce certain amount of so-called "3D noise", miscalculated 3D artifacts that occur when the algorithm does not make the correct calculation of the depth information. There is also a possibility of missing parts of 3D information due to the inherent problems that acquisition systems have to deal with (in case of computer vision system, for example, those would be specular features and flat surfaces without texture that might be present on the objects in the real world).

Synergetic mix of Computer Graphics and Computer Vision Systems

Our decision was to work on the system that would use the best of two worlds, a synergetic mix of computer graphics system and 3D acquisition systems. Instead of building elaborate models of the real world and then using processing power to animate and simulate those models in real time, we take an approach supported in 3D acquisition systems - we sample, transmit, and reconstruct the real world. We apply this technique of sampling not only to the participants in collaborative session but to every object in the environment whose existence as a "real" object is meaningful for a particular task. Our hypothesis is that the realistic representation of participants will have an effect on sense of presence; realism will also heighten our trust and engagement in communication, which in turn might affects our task performance. In some applications, such as tele-diagnosis, the representation of the real world obtained from the acquisition systems will be the only way to communicate the information (the doctor has to see what the real patient looks like).

Among the possible options for 3D acquisition systems, we decided to use a system harmless to the normal acting of the humans. The second requirement was to deploy the system that would be using off the shelf components. Our choice were passive methods i.e. computer vision systems that use a set of 2D cameras to “scan” the real world [9]. Starting with multiple 2D images of real objects, the typical vision system has to resolve the 3D geometry of the objects and their surface properties (texture).

Image 1: Graduate student at UNC Chapel Hill collaborates with remote colleague situated in Armonk (the prototype system demonstrated in October 2000)

To enrich the expressiveness of human communication, our prototype system supports the creation and direct manipulation of synthetic (virtual) computer graphics objects. Those synthetic objects coexist with "real" objects, providing "a best of both worlds" hybrid. Merely producing a close replica of what already exists in the real world is necessary, but not sufficient. An effective tele-immersive environment will empower us to do things that we would not be able to do in any other way. Those synthetic objects are usually the very subjects of human collaboration. Their existence might also be essential for the communication flow and the task that particular application is built around. Image 1 shows 3D acquisition data of remote participant being displayed locally together with a miniature synthetic model of an office design seen and manipulated by both participants. Another application example would be a group of orthopedic surgeons using different graphics modeling tools to design a prosthesis for the patient whose 3D representation might be a part of the same environment.

Proof-of-concept Tele-immersion Portal

The vision that the Initiative had in 1997 (Image 2, left) was a system that would be situated in the corner of one's office, where projectors would be displaying stereo images on two adjoining walls appearing as transparent passages to someone else's offices. 3D environment was to consist of a mix of 3D real-time acquisition data representing remote participants, and synthetic objects. As we mentioned earlier, this would represent a meaningful mix of "real" and "virtual" objects. The participants had to be able to use interactive graphics interfaces (virtual laser, for example) to point and manipulate with synthetic objects. The selected display stereo technology was passive, polarized stereo requiring the use of polarized glasses. The glasses look very much like ordinary sun glasses; being very light, with wide field of view and no connecting cable, it all minimizes the discomfort usually present when using active stereo and shutter glasses for example.

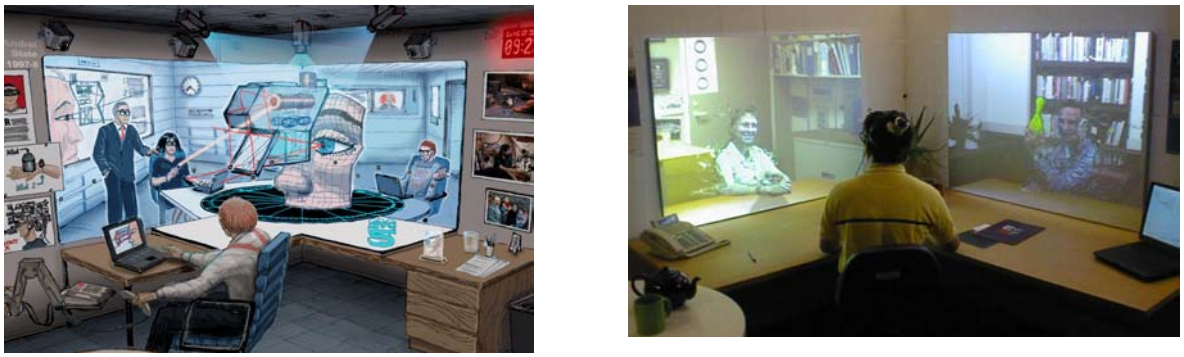

Image 2: vision from 1997 on the left (sketch done by Andrei State) and a version of prototype system from May 2000 on the right

In May 2000 (Image 2, right), and October 2000 (Image 1), we demonstrated proof-of-concept tele-immersive portal with all the visual features of the system as envisaged in 1997. While working on the system we were faced with challenges in almost every aspect of the design of our prototype, including: 3D dynamic (real-time) and static scene acquisition, dynamic scene modeling algorithms, rendering, high precision tracking, stereo projective displays, multi-person interaction, easy interoperation of heterogeneous stand-alone 3D graphics applications, and networking issues.

Our goal was to design a tele-immersion portal that allows tele-collaboration between multiple users without leaving their offices. It led us towards defining an arena that allows people to sit comfortably at their desks, move and make typical gestures for people engaged in a dialog with someone who might be sitting on the opposite side of their desk. This represents a working volume of reasonable size (1m x 1m x 0.5m), in which we can perform 3D acquisition of a real environment. In order to cover this working area we use a set of seven digital cameras positioned in a semi-circle and directed towards the area we are interested in. The 2D images are acquired by these cameras, and they are distributed to a group of five commodity PCs (quad Pentium III, 550MHz) that reconstruct 3D data ("confetti"). The 3D data are sent via Internet2 network link to a remote site where the final image is processed and displayed using a passive polarization stereo technique. Currently, the system is able to produce 2-3 such sets of 3D data (frames) every second. Maximizing frame rate was not the primary objective of our proof-of-concept prototype although it will be addressed in the future.

Even with a limited frame rate, our system achieves a dramatically heightened sense of presence compared to 2D video conferencing. We believe that this derives from the real-time proprioceptive feedback (matching between body movements and images on the visual display). Shifting one's head in tele-immersion system causes the appropriate change in the image of the remote environment and reinforces presence. Not only is it possible to have true eye contact, but you can also have gaze awareness, which is important for human communication. The subjects who have already experienced the system have usually forgiven visual noise (miscalculated 3D points) and concentrated on the fact that they could see the remote person in stereo and life size as if they were sitting on the opposite side of their desk. Being able to manipulate 3D synthetic objects (creation, selection, modification) using a virtual laser pointer coupled to a physical laser pointer provides an addition mechanism that strengthens the sense of presence. Although these preliminary results are encouraging, appropriate user studies should be taken to quantify the effects.

Applications and Future Vision

The tele-immersive portal could be characterized as a telephone of the future, a kind of interactive user interface that could be employed in a number of applications that use high bandwidth networks as a platform. Typical applications that will benefit greatly

include organizing tele-meetings, a variety of medical applications (preoperative planning, tele-assisted surgery, advanced surgical training, tele-diagnostics), tele-collaborative design, computer supported training and education, true 3D interactive video and most certainly entertainment.

Our long term goal is to provide a foundation for truly interactive and collaborative environments. However, much work remains in the visual part of the system. The important issues include: making 3D acquisition faster and the 3D data set cleaner (removing the 3D "noise"), providing tools that allow seamless and intuitive interaction with synthetic objects, and making the entire solution as compact as possible. We also plan to integrate a novel concept we have been working on and which has been developed and tested between different computer graphics applications. The concept is called SGAB (Scene Graph As Bus) and it provides the way in which two heterogeneous, stand-alone applications can dynamically exchange their scene graphs without significant modification of the code. In this way, exchanging the content the applications will exchange the functionalities embedded in each application. This enables free collaboration in computer-graphics client communities; it provides a way in which remote participants would be able to use interactive graphics tools that exist in other applications and their use could be meaningful in context of the local application [1].

Future extensions of the system will incorporate new sensory information such as spatialized audio, improved haptics and new interaction devices that provide not one but set of points of presence (ideally a surface), autostereoscopic displays, smart-camera clusters where cameras are used for both 3D acquisition and head and object tracking, and semantic parsing of the acquisitied scene and ability to distinguish individual objects within the 3D scans sent from remote sites.

The preliminary steps of our research make us increasingly optimistic that in the future this technology will provide highly compelling, immersive environments for an increasing variety of tele-collaborative applications over high bandwidth networks.

Acknowledgements

We would like to thank our sponsor and partner “Advanced Network & Services” for starting and supporting Initiative, as well as National Science Foundation for additional sponsoring of individual university partners. We would also like to extend special thanks to a number of researchers and graduate students from partner sites that took part in the Initiative during last four years.

References

[1] Zeleznik, B., Holden., L., Capps, M., Abrams, H., and Miller, T., "Scene-Graph-As-Bus: Collaboration Between Heterogeneous Stand-alone 3-D Graphical Applications", Eurographics 2000

[2] Sutherland, I.E., "The Ultimate Display", Proceedings of IFIP 65, Vol 2, pp. 506-508, 582-583, 1965.

[3] Ramesh, R., Welch, G., and Fuchs, H., "Seamless Projection Overlaps using Image Warping and Intensity Blending", 4th Annual Conference on Visual Systems and Multimedia, November 18-20, 1998, Gifu, Japan.

[4] Slater, M., and Wilbur S., "A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments", Presence: Teleoperators and Virtual Environments, Vol. 6, No. 6, 1997, pp. 603-616, MIT Press.

[5] Durlach, N., and Slater, M., "Presence in shared virtual environments and virtual togetherness", Proceedings of the Presence in Shared Virtual Environments Workshop, First International Workshop on Presence, Ipswich, Suffolk, UK, 1998.

[6] Tromp, J., Steed, A., Frecon, E. Bullock, A., Sadagic, A., and Slater, M., "Small Group Behaviour in the COVEN Project", IEEE Computer Graphics and Applications, Vol. 18, No. 6, pp. 53-63, 1998.

[7] Karla, P., Magnenat-Thalmann, N., Moccozet, L., Sannier, G., Aubel, A., and Thalmann, D., "Real-time Animation of RealisticVirtual Humans", IEEE Computer Graphics and Applications, Vol.18, No. 5, 1998, pp.42-55.

[8] Lee, W.-S., and Magnenat-Thalmann, N., "Fast Head Modeling for Animation", Journal Image and Vision Computing, Volume 18, Number 4, 1 March, 2000, pp.355-364, Elsevier.

[9] Mulligan, J., and Daniilidis, K., "View-independent Scene Acquisition for Tele-Presence", In Proceedings Int. Symposium on Augmented Reality, Munich, Germany, 2000.