Building a Panoramic Recording and Presentation System for TelePresence

Dávid Hanák, Gábor Szijártó, Alex Beregszászi, Gergely Mészáros Komáromi, Barnabás Takács

MTA SZTAKI Virtual Human Interface Group, Budapest, Hungary / BTakacs@sztaki.hu

Abstract

We present herein a panoramic capture and transmission system for the delivery of Internet-based telepresence services. Our solution involves a compact real-time spherical video recording setup that compresses and transmits data from six digital video cameras to a central host computer which in turn distributes the recorded information among multiple render- and streaming servers for personalized viewing over the Internet or 3G mobile networks. Our architecture offers a low-cost and economical alternative for personalized content management and it can serve as a unified basis for novel applications.

Keywords: PanoCAST, Telepresence, Immersive Spherical Video, Internet-based broadcast architecture

1. Introduction

Telepresence and remote operators that employ virtual-reality for education and entertainment have long been explored by scientists and developers of complex systems alike. The word *telepresence* is defined as "the experience of or impression of being present at a location remote from one's own immediate environment" [1]. To achieve a high level of immersion, a number of sensory stimuli, such as visual, auditory, tactile, and perhaps olfactory, need to be captured, encoded, transmitted and subsequently presented or rendered to the user in a real-time and fully transparent manner. Of course, the first step in building such technical solutions requires the availability of sensors that can capture and re-transmit relevant data and output devices that can render the information with as minimal distortion as possible. While the level of immersion in a telepresence system may be affected by many variables and measured with the help of *Presence Questionnaires* [2], the ultimate goal of such technical solutions, still remains the purpose to provide their users with the most up-to date information and control over a remote environment. In our current research therefore we focused on presenting visual and auditory stimuli only, but doing so to multiple viewers at a time and allowing them to share their experience. Video-based telepresence solutions that employ panoramic recording systems have recently become an important field of research mostly deployed in security and surveillance applications. Such architectures frequently employ expensive multiple-head camera hardware and record data to a set of digital tape recorders from which surround images are stitched together in a tedious process [3]. These cameras are also somewhat large and difficult to use and do not provide full spherical video (only cylindrical), a feature required by many new applications. More recently new advances in CCD resolution and compression technology have created the opportunity to design and build cameras that can capture and transmit almost complete spherical video images [4,5], but these solutions are rather expensive and can stream images only to a *single viewer*.

Gross et al. [6] describe a telepresence system called *blue-c*, which, using a CAVE system, a set of 3D cameras and semi-transparent glass projection screens, can create the impression of total immersion for design and collaboration. This system, however, requires expensive equipment and a complicated setup, therefore it is not feasible for servicing masses simultaneously (permitting it was not designed wih this goal in mind either).

Rhee et al. [7] present a low cost alternative to the above system, by adding cheap video cameras and sophisticated imaging algorithms to an existing CAVE system. However, the focus is still on collaboration between a limited number of participants.

A number of researches [8-10] target telepresence for robotic surgery, but again due to the different requirements, these systems are cleary not applicable in mass broadcasting.

To address the above difficulties we have developed a broadcasting solution, called *PanoCAST* that is capable of recording and simultaneously streaming live 360 degree full spherical video images to remote users over digital networks, such as the Internet or 3G mobile phones, while allowing them to control their own point of view with the help of virtual cameras.

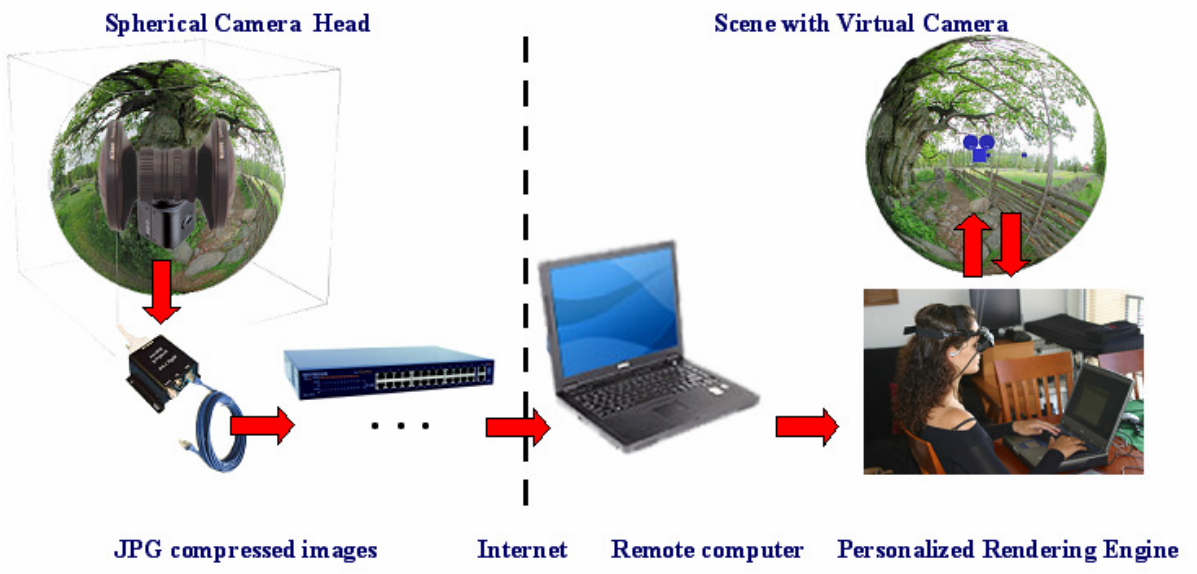

Figure 1: *Functional overview of our telepresence system.*

The remaining of this paper is organized as follows: In Section 2 we review the overall architecture of our system. In Section 3 we present details on its implementation, while Section 4 contains our conclusion and discusses future work.

2. PanoCAST System Architecture

To record and stream high fidelity spherical video we employ a special camera system with six lenses packed into a tiny head-unit. The images captured by the camera head are compressed and sent to our server computer in real-time delivering up to 30 frames per second, where they are mapped onto a corresponding sphere for visualization. The *PanoCAST* system then employs a number of virtual cameras and assigns them to each user who logs-in over the Internet, thereby creating their own, personal view of the events the camera is seeing or has recorded. The motion of the virtual cameras is controllable via TCP/IP with the help of a script interface or can be directly controlled by physical sensor data encoding the head motion (e.g. the output of an orientation tracker attached to a head mounted display) of the user on the remote site. The resulting images the users each see can be streamed to their location using RTSP protocol for mobile devices or video-conferencing tools, such as Skype via a special client-server solution. Finally, on the client side, the system can accommodate a variety of different displays and input devices, including a HMD where the head motion of the user directly controls the rotation of the virtual camera, thereby delivering a sensation of presence.

Figure 1 shows the basic setup and our approach to telepresence for demonstrated for a single user. A spherical camera head (left) is placed at the remote site in an event where the user wishes to participate. The camera head captures the entire spherical surroundings of the camera with resolutions up to 3K by 1.5K pixels and adjustable frame rates of maximum 30 frames per second (fps). These images are compressed in real-time and transmitted to a remote computer over G-bit Ethernet connection or using the Internet, which decompresses the data stream and remaps the spherical imagery onto the surface of a sphere locally. Finally, the personalized rendering engine of the viewer creates TV-like imagery and sends it to a *Head Mounted Display* (HMD) with the help of a *virtual camera* the motion of which is directly controlled by the head turns of the user.

In principle, for multiple viewers, this simple architecture can be easily modified to accommodate a number of independent HMD devices each controlling their own respective virtual camera. The technical difficulty in creating such system, however, lies in the bandwidth required to distribute the high quality images directly to each user while it also lays large computational burden on the local computer.

The key idea behind our solution is based on distributing each user only what they currently should see instead of the entire scene they may be experiencing. While this reduces the computational needs on the receiver side (essentially needing only to decode streamed video and audio data) and send tracking information and camera control back in return, it

places designers of the server architecture in a difficult position.

To overcome the limitations we devised an architecture shown as a box diagram in Figure 2. The *panoramic camera head* is connected via an optical cable to a *JPG compression* module, which transmits compressed image frames at video rates to a distributions server using IEEE firewire standard. The role of the *distribution server* is to multiple the data video data and prepare it for broadcast via a server farm. To maximize bus capacity and minimize synchronization problems, the distribution server broadcasts its imagery via *UDP protocol* to a number of *virtual camera servers*, each being responsible for a number of individually controlled cameras. The number of these server computers is governed by the number of clients the system needs to service in parallel at any-given moment. Their role is to compute user-dependent virtual views of the panoramic scenery using one camera for each connected user or a group of users who control what they see in competition with one another. With hardware acceleration incorporated in modern graphics cards or GPUs, a single unit can service up to m=20 independent camera views. Video data is then first encoded in MPEG format and subsequently distributed among a number of *streaming servers* using RTSP (Real-time Streaming Protocol) before sent out to individual clients over the Internet or 3G mobile networks. Assuming 3Gbit/sec connection a streaming server is capable of servicing up to 100 simultaneous clients at any given moment. Again, the number of streaming servers can be scaled according to the need of the broadcast. Finally, independent image streams are synchronized with audio and arrive at the user site ready to be decoded and displayed.

In the *PanoCAST* telepresence system interaction primarily means that the user controls the orientation and field of view of the camera while observing a remote scene or event taking place. This functionality is implemented via a script-based command

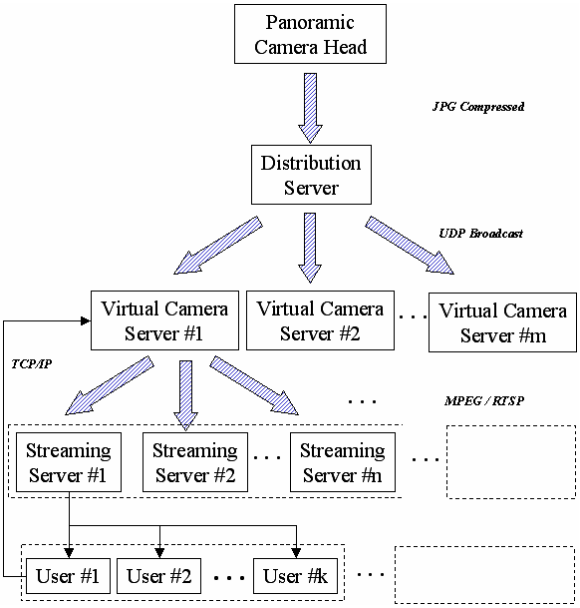

Figure 2: *Server-park and data flow architecture for independently controlled viewer experience.*

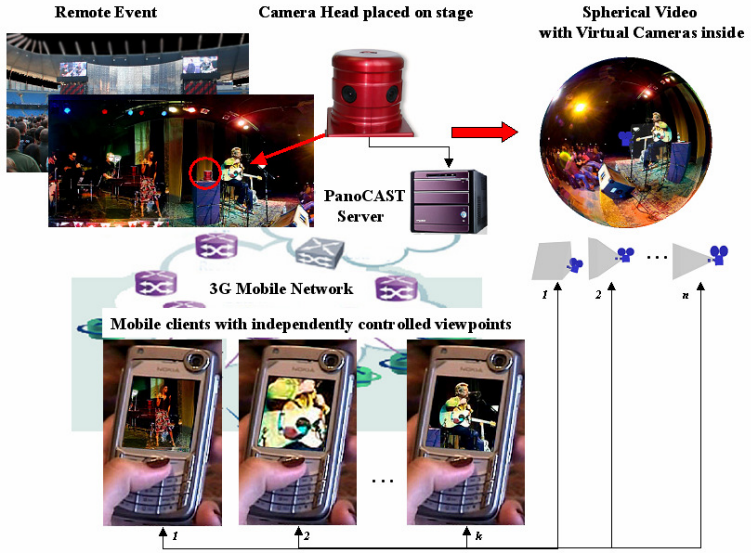

Figure 3: *Example of using PanoCAST telepresence technology to stream personally controlled independent views to mobile phones over 3G networks.*

interface that sends either discrete commands to rotate the camera in a certain direction (e.g. when controlled from a web-browser) or using continuously varying physical device data such as a head tracker, mouse, the output of facial analysis software or simple game controllers (see below). This interaction takes place via *TCP/IP protocol*. As each viewer is allowed to control their own camera or join a group of people viewing the same portion of reality, the resulting experience is as if he or she was present. Similarly, when the end point of video streaming is a mobile phone with 3G connection, the *PanoCAST* solution offers a unique point of view and entertainment value. This is demonstrated in Figure 3 for a live music concert situation. In the following section we discuss some of the key technical elements of our solution on more detail.

3. Implementation Notes

One of the key elements of the *PanoCAST* system is the compact and portable 360 degree panoramic video recording system depicted in Figure 4. It was designed to minimally interfere with the scene being recorded. Since almost the entire spherical surroundings are recorded working with such a camera is rather difficult from a production's point of view. Specifically, the basic rules and the concept of frames here become obsolete, as both lighting, microphones as well as the staff remains visible. To provide as much immersion as possible, the camera head is placed on the end of a long pole carried by the cameraman (in this case a camerawoman shown). This setup is similar as if we replaced a person's head standing at a given location with the camera or in other words, it is the ultimate steady-cam where the viewer may almost directly participate in the chain of events. By virtue of the extended fixture, it is possible to look around, even under ones feet or "up to the skies" without disturbance from the mounting structure.

The computer to control the capture process is located in the mobile rack shown on Figure 5. The heart of the system (see left) is a small-factor personal computer (Apple MacMini) which is controlled via a touch screen interface or occasionally standard keyboard during a recording session. Video is digitally stored on an external drive recording up to 1.5 hours of video on a 250Gbyte SATA unit. The continuous power supply that allows for 1.5 hours of operation can be seen below. On the right the same rack is shown while it is worn by another member of the staff.

To enhance the functionality of our broadcasting solution, we have enabled our server-architecture

Figure 4: *Portable 360^0 panoramic camera head.*

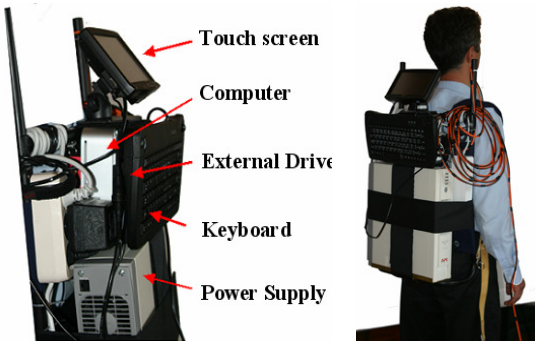

Figure 5: *Portable PanoCAST recording system.*

with multiple receivers on the client side. The first possibility to receive *PanoCAST* video via the Internet uses a *virtual camera driver* that allows any application to receive video from the camera server as if it was a simple web camera connected to the computer. In fact, the operating system sees these devices and handles them much the same way as it is doing with physical devices. This is shown in Figure 6 where on the left side the Windows device manager shows four virtual cameras installed, each receiving its input from a different render unit, and outputting its content to any video-based communication application, such as *Skype* (shown right), *Yahoo Messenger* or *Microsoft Messenger*. The second output option is using *MS WMV broadcasting*, whereas any *Media player* on the client side may connect to a data stream and observe the output of the cameras (Figure 7 left).

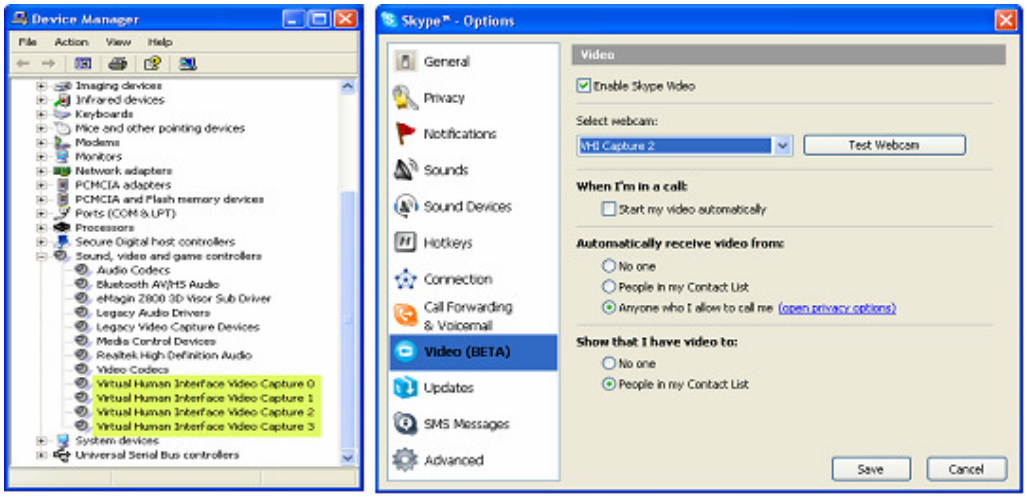

Figure 6: *Virtual camera drivers installed in a system (left) and a videoconferencing application (Skype) using one these cameras instead of a web camera.*

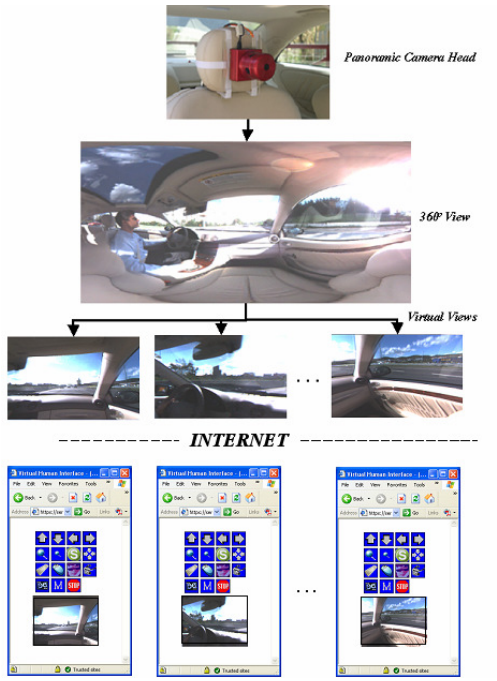

Figure 7: *PanoCAST dataflow with Internet-based viewers on the client side.*

Finally, for individual camera control, the streaming servers in the architecture allow each client to receive personalized video content in a browser using *Flash* or our own *ActiveX* controller as demonstrated in Figure 7. The figure shows the final data flow of the *PanoCAST* architecture with the Internet-based viewers shown in the bottom row. These browser-based viewers on the client side allow several different ways for the user to control the rotation and field of view parameters of their respective virtual cameras. This is the subject of the remaining of this section.

Interactive camera control in the proposed telepresence system occurs via direct input from the user, e.g. keyboard strokes or from sensory information obtained from physical devices, such as a head tracker, mouse or game controller. The head tracker interface obtains yaw, pitch and roll parameters from the HMD and translates those into camera rotations by sending them over the TCP/IP connection to the host application. When the delay in the digital network is minimized, this leads to an interactive experience similar to being present in the VR room. Similarly, mouse information is mapped from screen space onto rotations of the virtual cameras while a similar solution exists for using game controllers (most notable the *Wii by Nintendo*) that provide intuitive control. Finally, the face detection capabilities of the VHI architecture also allow the viewer to look around by simple moving his or her head in front of the computer screen.

4. Conclusion and Future Work

In this paper we have introduced a multi-cast application capable of real-time streaming and control of spherical video images over digital networks for multiple viewers sharing the same experience, but from different perspective. Using this architecture we developed intuitive user controls and multiple digital network interfaces that allow for creating a number of novel applications that involve telepresence. Specifically, our system, called *PanoCAST,* has been tested in a number of digital networks including wired-Internet, WIFI and 3G solutions. Test results showed that a single server computer can deliver

services to up to 20 clients with reasonable delays. Several test production videos have been recorded demonstrating the applicability of our solution, while the system is currently being deployed for commercial use. We argue that such a technical solution represents a novel opportunity for creating compelling and content for the purposes of education, entertainment and many other application areas.

5. Acknowledgment

The research described in this paper was partly supported by the *PanoCAST Corporation*, Budapest, Hungary (http://www.PanoCAST.net) and the *VirMED Corporation*, Budapest, Hungary (http://www.VirMED.net).

References

[1] Transparent Telepresence Research Group (2007), http://www.telepresence.strath.ac.uk/telepresence.htm

[2] Witmer, B.G., M.J. Singer, (1998), "Measuring Presence in Virtual Environments", in Presence, **7** (3): pp.225-240.

[3] Pryor, L., A.S. Rizzo (2000) "User Directed News" *http://imsc.usc.edu/research/project/udn/udn_nsf.pdf*

[4] Immersive Media Dodeca Camera (2007), http://www.immersive-video.eu/en

[5] Point Grey Research, LadyBug2 Camera (2007), http://www.ptgrey.com/products/ladybug2/index.asp

[6] Gross, M., Würmlin, S., et al. (2003), "blue-c: A Spatially Immersive Display and 3D Video Portal for Telepresence", in ACM Transactions on Graphics, Vol. **22 #3** pp. 819-827.

[7] Rhee, S.M., Ziegler, R. et al. (2007), "Low-Cost Telepresence for Collaborative Virtual Environments", in IEEE Trans Vis Comput Graph, Vol. **13 #1** pp. 156-166.

[8] Ballantyne, GH (2002) "Robotic surgery, telerobotic surgery, telepresence, and telementoring", in Surgical Endoscopy, Vol. **16 #10** pp. 1389-1402.

[9] Latifi, R. Peck, K. et al. (2004) "Telepresence and telementoring in surgery", in Stud Health Technol Inform, Vol. **104** pp. 200-206.

[10] Anvari, M. (2004) "Robot-assisted remote telepresence surgery", in Semin Laparosc Surg, Vol. **11 #2** pp. 123-128.