Interactive 3D Teleconferencing with User-adaptive Views

Ruigang Yang
Department of Computer Science
University of Kentucky
Lexington, Kentucky U.S.A.
ryang@cs.uky.edu

Andrew Nashel and Herman Towles
Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina U.S.A.
(nashel,herman)@cs.unc.edu

ABSTRACT
We present a system and techniques for synthesizing views for three-dimensional video teleconferencing. Instead of performing complex 3D scene acquisition, we decided to trade storage/hardware for computation, i.e., using more cameras. While it is expensive to directly capture a scene from all possible viewpoints, we observed that the participants’ viewpoints usually remain at a constant height (eye level) during video teleconferencing. Therefore we can restrict the possible viewpoint to be within a virtual plane without sacrificing much of the realism. Doing so significantly reduces the number of cameras required. We demonstrate a real-time system that uses a linear array of cameras to perform Light-Field style rendering. The simplicity and robustness of light field rendering, combined with the natural restrictions of limited view volume in video teleconferencing, allow us to synthesize photo-realistic views per user request at interactive rate.

Categories and Subject Descriptors
H.5.1 [Multimedia Information System]: Video teleconferencing

General Terms
Design, Algorithms, System

Keywords
3D video-teleconferencing, Light field rendering, View synthesis

1. INTRODUCTION
With recent rapid advances in network bandwidth and dropping costs for video equipment, video teleconferencing, a technology enabling communicating with people face-to-face over remote distances, has been widely deployed for business and education. Currently, the majority of video teleconferencing applications use a single camera to capture the participant(s) from a fixed viewpoint and directly display the video stream in the remote site. While widely used, this design suffers from a number of drawbacks, such as low resolution, fixed point of view, and limited field of view. Therefore it does not provide a compelling or convincing presence to the participants [5].

The Office of the Future group at the University of North Carolina at Chapel Hill in 1998 introduced a vision for the ultimate tele-conferencing/collaboration interface [4]. In their long-term vision, an ordinary office is equipped with “a sea of cameras” and projectors [2]. The complete, dynamic 3D scene is extracted using computer vision techniques and transmitted over the network to a remote office. A unique view is then rendered in life size for each remote viewer. Thus collaborators in any locale would be able to interact with each other as if they were in a common room. The challenges in implementing this interface are enormous, particularly in 3D scene reconstruction. This problem has been one of the central topics for computer vision for decades. While many algorithms exist, they are quite fragile in practice [3].

In a previous paper [6] we presented an alternative design that allows view synthesis without 3D reconstruction. This demonstration, based on our paper, shows a prototype...
for live photo-realistic 3D video teleconferencing over the internet. For the sake of completeness, we briefly overview the underlying techniques here, details can be found in the original paper [6].

Our system is based on the Light Field Rendering (LFR) technique [3]. The basic idea of LFR is to use many cameras to record the flow of light in all directions, thus the task of view synthesis becomes a simple table-lookup of the view rays, bypassing the difficult 3D reconstruction problem. LFR typically requires hundreds or even thousands of cameras to cover the scene. For our teleconferencing setup, we observed that during a video teleconferencing session, the participant’s view point is quite limited, usually at the eye level, with small lateral motions. Thus we can use a 1D linear array of cameras to capture a compact light field, which we refer to as the Line Light Field. This compact 1D setup makes real-time capture, transmission, and rendering possible. Novel views at eye level can be changed interactively, allowing the participants to view the remote scene from side to side, or from near to far to gain a sense of 3D. Furthermore, we can synthesize large FOV images using cameras that do not share a common center of projection.

2. SYSTEM DESCRIPTION

We have implemented a 3D teleconferencing prototype between the University of Kentucky (UKy) and the University of North Carolina at Chapel Hill (UNC-CH). Each site has a total of eight Sony digital firewire cameras arranged in a linear array, as shown in Figure 1. These cameras are regularly placed at 65 millimeter apart, very close to the minimum distance allowed by the form factor of the camera body. All cameras are synchronized by a wire controlled from a PC and fully calibrated using the method from [7].

The video acquisition system (one in each site) includes four PCs interconnected through 100Mbit Ethernet. Each of them is connected to two Sony cameras and is used to capture and JPEG-encode the raw image data at full VGA resolution. The JPEG streams can be sent through the network.

We use light field rendering [3, 1] to synthesize new images from a user-driven virtual camera. There are two options to run the view synthesis program. One is to send all the video streams over the internet and synthesize novel views at the remote site (remote rendering). Alternatively we can synthesize views locally and only send the final result to the remote site (local rendering). The first approach has a lower latency for changing viewpoint while the second is easier to manage from a network standpoint. In terms of scalability, the bandwidth requirement for remote rendering is \(O(kn)\) where \(k\) is the number of cameras and \(n\) is the number of sites (assuming multicast is used). The bandwidth requirement for local rendering, on the other hand, has a quadratic growth rate at \(O(n^2)\). Given that our current prototype includes only two sites and the viewpoint is not changing rapidly (i.e., no head tracking), we chose to implement the local rendering approach.

Locally, we can achieve an update rate of 4-7 frames per second (fps) for VGA input images. The bottleneck is in image capture. We can only capture synchronized VGA resolution images at 7-8 FPS with two cameras on the same 1394 bus. This is caused by the 1394 bus bandwidth limitation and the transfer characteristics of the digital cameras under external trigging. The synthesized view, typically at \(1024 \times 512\), is read back from the rendering program’s framebuffer and sent to the remote site through TCP/IP with JPEG encoding. The frame rate between UKy and UNC-CH varies from 5 fps to 10 fps, depending on the network traffic. Optimization in the network code or the use of a more sophisticated compression scheme is expected to substantially increase the frame rate.

Our demonstration will include one live 3D video stream from UKy or UNC-CH.

3. ACKNOWLEDGMENTS

The authors would like to thank Xinyu Huang, Celso Kurashima, Anselmo Lastra, and Henry Fuchs for their participation in this project. This research was funded in part by the office of research at the University of Kentucky and Sandia National Laboratories under the Department of Energy’s ASC VIEWS program B519834.

4. REFERENCES


