A Test-Bed for Fly-Around 3-Dimensional Television

Category: Research Paper, Online ID: Papers_356

Abstract

This paper describes a system for simulating the experience of viewing a light field television of the future. The system uses a camera array employing 112 imagers. Head tracking is combined with stereo display to give a foretaste of a true fly-around 3-dimensional television system. Such a system would broadcast a continuous sweep of orientations of a scene and would be viewed with a display capable of presenting the scene with correct stereo and motion parallax cues. This paper presents implementation details and lessons learned from the fly-around camera. It also describes preliminary results that suggest promising types of content for this new medium.

Keywords

Computer Vision, Graphics Hardware, Graphics Systems, HCI (Human-Computer Interface), Hardware Systems, Human Visual Perception, Image-Based Rendering, Multimedia, Optics, Video, Virtual Reality, Head Mounted Displays.

1.0 Introduction

Image-based methods for representing 3-dimensional scenes have become increasingly popular in recent years because of their ease of capture and generality of use. Schemes may be purely image-based such as the light fields of Levoy and Hanrahan [Levoy1996] or use approximate surface hulls along with dense ray samples such as lumigraphs [Gortler1996].

By combining light field rendering with low-latency head-tracking and stereo display, it has proved possible to create convincing illusions of static 3-dimensional objects [Regan1999] using a form of "Fish Tank" virtual reality [Deering1992]. At the same time, progress has been made towards capturing multiple viewpoints of dynamic scenes using arrays of cameras. [Deering1994] used 32 cameras to view medium angular-resolution light fields for remote telepresence applications. [Dodgson1997] combined 12 video streams with a novel autostereoscopic display. In a similar approach to our own, [Kunita1999] combined signals from 12 genlocked video cameras to allow for live remote viewing. [Rander1997] used an array of cameras over the surface of a sphere to capture motions within a volume. Because of the sparse sampling of orientation space, it was necessary to use vision algorithms to reconstruct 3-dimensional models to generate intermediate views. While this approach proved successful for a variety of somewhat smooth, mostly diffuse objects, such as people performing sports. It is an open question as to the quality with which such algorithms can represent arbitrary structures and materials like sparkling gems, layered glass, flower arrangements etc. In contrast, [Miller1999] used a motion gantry to capture time-lapse light field movies of slowly changing scenes with between 100 and 256 orientations per sweep, allowing the representation of intricate structures and optical properties. The results of these experiments were sufficiently compelling that it is tempting to imagine a future video standard based upon the motion-light-field or fly-around-television concept.

Section 2 of this paper describes scenarios for the deployment of incrementally more complete versions of this idea, listing the required technical infrastructure and restrictions on viewing experience. Section 3 describes the construction and use of a live light-field video camera with 112 sensors. Section 4 summarizes the results of content and interaction experiments with the system and suggests the next steps to take in perfecting this new medium.

2.0 Steps towards Fly-Around 3-D Television

This section describes ways in which a traditional video stream may be enhanced to help portray a scene in a way that is more immersive and engrossing.

2.1 Multi-viewpoint Video

One simple way to enhance coverage of a scene is to include several camera angles with a way to switch between them such as a remote control. Such ideas have been incorporated into digital video discs and coverage of sporting events on satellite television. If the cameras face in towards a central region while placed along a circular arc then we have a system for "fly-around video". As the number of cameras is increased, images from adjacent ones may be blended to get intermediate images since the samples approach the density of a light field. As with light fields, there is a working range of depths within which the images can be blended successfully [Kunita1999].

2.2 Head-Tracked Light-Field Video (single user)

One way to interact with a fly-around video stream is to use a remote control to determine the camera from which the scene is viewed. This has the advantage of being relatively tolerant of large angular jumps between adjacent views and was the basis of "virtual object

movies" in QuickTimeVR [Chen1995]. A second approach uses head tracking to update the viewing projection. Since the scene must change smoothly in response to head motion, this approach requires higher angular resolution for the light field stream. Also, when combined with a conventional stereo display, only one user at a time perceives the correct view. However, the illusion of a stable 3-dimensional scene is more compelling since the inertial cues of one's own motion combine with the visual display to create a more convincing illusion of looking at a real object. An important experiment is to determine what sort of content is most enhanced by head-tracked fly-around video versus conventional video. Our test-bed was designed to help answer this question as well as to motivate further technical refinements if the user experiences were found to be compelling.

2.3 Light Field Displays (multi-user)

Head-tracked stereo displays are best suited to single users in front of a personal display (such as for a personal computer). To allow a multi-viewer experience either requires increasingly taxing time multiplexing with accurate tracking or a display capable of showing a light field directly, with the appropriate image being presented in the corresponding direction. Such displays may use holography [St. Hilaire1998] or arrays of projectors with special screens [Borner1993] or time-multiplexed displays with steerable optical elements [Dodgson1997]. A disadvantage of such displays is that the whole light field stream must be decompressed and presented, and scaling such systems up to hundreds of orientations (to enable adequate depth of field) will remain challenging for the foreseeable future. However, in the long term, such displays may prove the most convenient, avoiding tracking errors and the need for special glasses.

3.0 Live Light Field Camera Experiments

In order to study user interactions with real-time content we decided to build a live fly-around television camera. By making a live system, we obviated the need for large amounts of storage as well as facilitating spontaneous interactive experiments.

Our system consisted of 112 genlocked CMOS video cameras arranged over a 79-degree arc with a radius of 9.5 feet. The camera is shown in Fig. 1. Previous experiments with gantry-based light fields had highlighted the need for this angular spread and resolution, to display objects of interesting depth and to enable user interactions with significant head motions [Miller1999]. This wide range of horizontal orientations, combined with a large number of imagers, is one of the novel aspects of our system over previous work. It enabled a less-confined study of user interactions and content.

In a similar design to [Kunita1999], an analog switching network was used to select video from any two adjacent cameras. These were then blended together using an analog cross-dissolve circuit shown in Fig. 2. The network was duplicated to work for two independent sets of cameras, allowing the generation of left and right eye images, each of which was blended from two adjacent cameras. These images were displayed as red-green stereo on a component video monitor, the user wearing red-green glasses.

(Note to reviewers: the system was used in monocular mode for the videotape, since red-green stereo transfers poorly to VHS tape. The visual effects are much more compelling in stereo.)

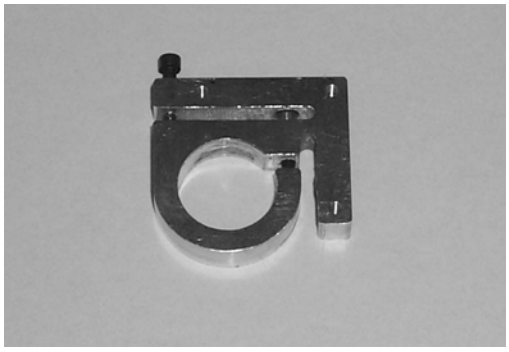
3.1 Camera Alignment

An original goal of the design was to allow for the selection of camera sources on a per-scanline basis. This then allows a more correct light-field resampling based on the fact that different portions of the screen are viewed from different orientations. The sources only need to change once per scanline if the system is arranged to have vertical scan-lines [Kunita1999] and only exhibits horizontal motion parallax. However, such a design assumes that the cameras are perfectly aligned since image warping is not possible, given that only the current scan-line is available from a given camera. Any image warping may require information from adjacent scanlines, which are not available.

Image correction can be done on a per-camera basis, with the resultant pixels being blended. However that would involve a frame-buffer and image warping hardware per camera - a prohibitive expense for the current project given the number of cameras. An alternative approach was taken using very precise mechanical alignment to remove as much distortion as possible from the sensors.

The cameras were evaluated for focal length and lateral displacement of the lens relative to the CMOS sensor. The lens was displaced relative to the sensor using a special jig with a calibrated target, and then locked in placed. Since we could not compensate for the field of view in the current design, we sorted the imagers by focal length, so that the imagers increased monotonically with location. The variation of the focal lengths was 3% over the range of cameras with an approximately linear progression from one end to the other. Since only adjacent cameras would be blended together this approach largely eliminated visual artifacts due to focal length variations.

To align the cameras in terms of orientation, we used a special flexure shown in Fig. 3. An array of mounted cameras is shown in Fig. 4. The flexure allows for gross alignment of camera twist by freeing the barrel-clamp and rotating the camera. Fine adjustments of the twist were made using the vertical screw adjustment. Pan and tilt of



the camera is controlled using compression springs combined with the three horizontal mounting screws.

Fig. 3. The camera flexure.

This cheap and compact flexure did allow orientation alignment to within a pixel given a calibration target. However, to our disappointment, the alignment showed drift over a few hours by as much as 3 pixels. We suspect that this may be the result of mechanical creep in the polycarbonate curved mounting plates and we hope to replace these soon with aluminum ones for the next iteration of the camera.

(The analog circuitry did allow compensation for variations in the gain of the cameras. However, there was not time before the submission deadline to calibrate and adjust these parameters which resulted in significant flicker in the video. We fully expect to be able to fix this for the final version of the paper.)

3.3 Head Tracking

Head tracking of a user's motion was achieved using an off-the-shelf USB digital camera modified with an infrared pass filter. The camera was augmented with four bright infrared light-emitting diodes, illuminating the viewer. The stereo glasses were coated with retroreflective material. This arrangement is shown in Fig. 5. The outline of the glasses shows up brightly against a dark background, and a simple tracking algorithm was able to find the location and orientation of the glasses at 60 frames per second. This location was used to switch between camera sources for left and right eye stereo during vertical sync for the camera. Fig. 6a and 6b show the display screen being viewed from oblique angles.

Due to tracking and display delay, the total system latency was between 16 and 30 ms. Previous work [Regan1999] suggested that latencies as low as 7 ms are detectable in an A/B comparison test. We found that subjects did not experience latency discomfort with our system, probably because motion within the scene masked the lag.

3.4 Content Experiments

Once working, the system was used to perform a number of informal content experiments. The subjective experience of using the system was that of looking at miniature puppets behaving like people or animals. Users tended to peer round objects to catch a view of otherwise hidden action and were altogether more physically engaged than with traditional video.

In one experiment, called "Marco Polo", an actor peered at a particular camera using a cardboard tube and called out "Marco". The viewer then positioned herself to peer directly down the end of the tube and called "Polo". This very simple activity was suprisingly fun and demonstrated the accuracy and rapidity with which the viewer could control her vantage point. In a second experiment, two actors assembled a flower arrangement in a vase, deliberately blocking the view of the flowers from certain vantage points. The resultant motions of the viewer were rapid and intimately related to the motion of the two actors. The viewer's motions enabled the flower arrangement to remain unoccluded and, at other times, provided motion parallax, which helped to reinforce the 3-D perception of the scene.

In general, user motion was more exaggerated and uneven than would be created by a production cameraman. However, this had the effect of heightening depth perception and the sense of immersion. When viewing certain subject matter, such as demonstrations of the assembly of complex objects, the lack of vertical motion parallax was sorely missed, highlighting the need for a true light field camera with a 2-dimensional array of imagers.

4.0 Conclusions and Future Work

While careful mechanical adjustment eliminated gross misalignment between the cameras, additional computational image rectification would be more precise and less tedious to set up. A cost-effective way to do this would be to abandon the requirement of switching camera sources once per scanline and to do it once per image. In that way, at the cost of some visual distortion, a single image correction circuit could be applied to each video channel at the end of the switcher. Alternatively, the imagers would need to be made random access rather than sequential.

A second improvement would be obtained using a digital switching network rather than the current analog one. Finally, a circuit needs to be added to perform the projective distortion needed to compensate for the screen being viewed off axis. In the longer term, of course, we would like a camera array that also records, to enable compression experiments and the production of more elaborate content.

Despite these limitations, the current design of the fly-around video camera, with over one hundred imagers, did

enable a variety of interesting experiments. Even relatively simple productions were found to be visually entertaining, and it was possible to see that this new visual medium would develop its own unique forms of presentation. Interacting with the displays using just head-tracked motion did prove physically tiring. As the system matures we hope to explore styles of interaction that combine explicit physical devices (such as remote controls) with head tracking.

13.0 References

[Borner1993]
Borner, R., "Autostereoscopic 3D-imaging by front and rear projection and on flat panels", Display, Vol. 14, No. 1, 1993.

[Chen1995]
Chen, Shenchang Eric, "Quicktime VR - An Image-Based Approach to Virtual Environment Navigation", Proceedings of SIGGRAPH 95, Computer Graphics Proceedings, Annual Conference Series, pp. 29-38 (August 1995, Los Angeles, California).

[Cutler1997]

Cutler, Lawrence D. , Bernd Frolich and Pat Hanrahan, "Two-Handed Direct Manipulation on the Responsive Workbench", 1997 Symposium on Interactive 3D Graphics, pp. "107—114, 1997.

[Deering1992]

Deering, Michael F., "High resolution virtual reality", Computer Graphics (SIGGRAPH '92 Proceedings), Vol. 26, 1992, pp. 195-202.

[Deering1994]

Deering, Michael F. "Facing the Challenge: Delivering Virtual Reality", Virtual Reality Software and Technology (Proceedings of VRST'94, August 23-26, 1994, Singapore), pp. 1-4 (August 1994, Singapore). World Scientific Publishing.

[Dodgson1997]

Dodgson, N. A., J. R. Moore and S. R. Lang, "Time-multiplexed autostereoscopic camera system", Proc. SPIE 3012, SPIE Symposium on Stereoscopic Displays an Applications VIII", San Jose, California, Feb 11-Feb 13, 1997.

[Gortler1996]

Gortler, Steven J., Radek Grzeszczuk, Richard Szeliski, Michael F. Cohen, “The Lumigraph”, Computer Graphics Proceedings, Annual Conference Series (SIGGRAPH 96), pp 43-54.

[Kunita1999]

Kunita, Yutaka, Masahiko Inami, Taro Maeda, Susumu Tachi, "Real-time Rendering System of Moving Objects", Proceedings of 1999 IEEE Workshop on Multiview Modelling and Analysis of Visual Scenes (MVIEW '99), pp. 81-88.

[Levoy1996]

Levoy, Marc, Pat Hanrahan, “Light Field Rendering”, Computer Graphics Proceedings, Annual Conference Series (SIGGRAPH 96), pp. 31-42.

[Miller1999]

Miller, Gavin S. P. , Steven M. Rubin, Philip M. Hubbard, Jenny Dana, and John Woodfill, "Head-Tracked Light Field Movies: A New Multimedia Data Type", Proceedings of Eurographics Multimedia '99 Workshop, Milan, Italy, Sept 7-8, 1999, pp. 141-152.

[Rander1997]

Rander, Peter W., P. J. Narayanan and Takeo Kanade. "Virtualized Reality: Constructing Time-Varying Virtual Worlds from Real World Events", IEEE Visualization '97, pp. 277-284 (November 1997). IEEE. Edited by Roni Yagel and Hans Hagen. ISBN 0-58113-011-2.

[Regan1999]

Regan, Matthew J. P., Gavin S. P. Miller, Steven M. Rubin and Chris Kogelnik, "A Real-Time Low-Latency Hardware Light-Field Renderer", Proceedings of SIGGRAPH 99, Computer Graphics Proceedings, Annual Conference Series, pp. 287-290 (August 1999, Los Angeles, California).

[St. Hilaire1998]

St. Hilaire, Pierre, "Holographic video: the ultimate display technology?", Optics and Photonics News. Vol. 8, No. 8 (August 1998), pp. 35-39.