3D TV: A Scalable System for Real-Time Acquisition, Transmission, and Autostereoscopic Display of Dynamic Scenes

Wojciech Matusik Hanspeter Pfister*

Mitsubishi Electric Research Laboratories, Cambridge, MA.

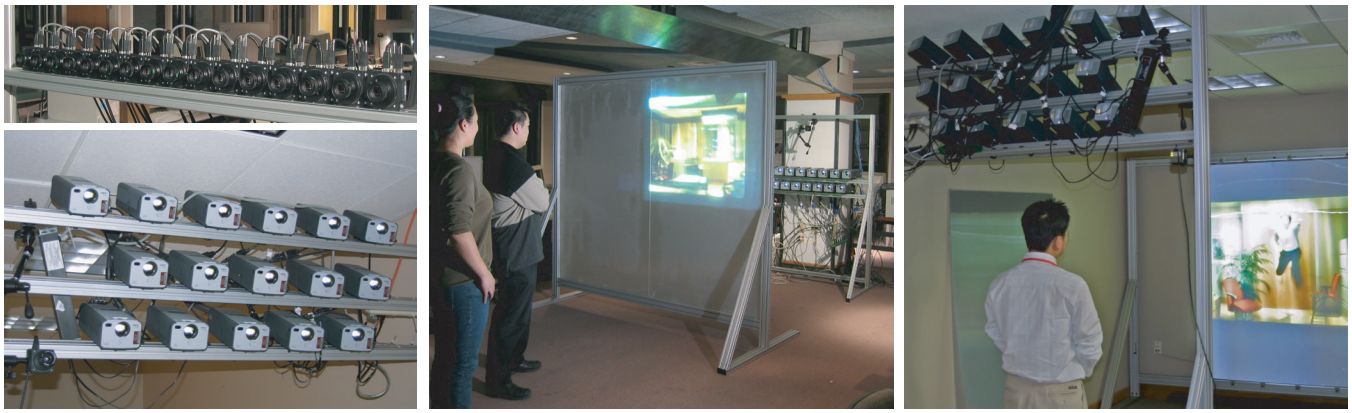

Figure 1: *3D TV system. Left (top to bottom): Array of 16 cameras and projectors. Middle: Rear-projection 3D display with double-lenticular screen. Right: Front-projection 3D display with single-lenticular screen.*

Abstract

Three-dimensional TV is expected to be the next revolution in the history of television. We implemented a 3D TV prototype system with real-time acquisition, transmission, and 3D display of dynamic scenes. We developed a distributed, scalable architecture to manage the high computation and bandwidth demands. Our system consists of an array of cameras, clusters of network-connected PCs, and a multi-projector 3D display. Multiple video streams are individually encoded and sent over a broadband network to the display. The 3D display shows high-resolution (1024×768) stereoscopic color images for multiple viewpoints without special glasses. We implemented systems with rear-projection and front-projection lenticular screens. In this paper, we provide a detailed overview of our 3D TV system, including an examination of design choices and tradeoffs. We present the calibration and image alignment procedures that are necessary to achieve good image quality. We present qualitative results and some early user feedback. We believe this is the first real-time end-to-end 3D TV system with enough views and resolution to provide a truly immersive 3D experience.

CR Categories: B.4.2 [Input/Output and Data Communications]: Input/Output Devices—Image Display

Keywords: Autostereoscopic displays, multiview displays, camera arrays, projector arrays, lightfields, image-based rendering

*[matusik,pfister]@merl.com

1 Introduction

Humans gain three-dimensional information from a variety of cues. Two of the most important ones are *binocular parallax*, scientifically studied by Wheatstone in 1838, and *motion parallax*, described by Helmholtz in 1866. Binocular parallax refers to seeing a different image of the same object with each eye, whereas motion parallax refers to seeing different images of an object when moving the head. Wheatstone was able to scientifically prove the link between parallax and depth perception using a steroscope – the world's first three-dimensional display device [Okoshi 1976]. Ever since, researchers have proposed and developed devices to stereoscopically display images. These three-dimensional displays hold tremendous potential for many applications in entertainment, information presentation, reconnaissance, tele-presence, medicine, visualization, remote manipulation, and art.

In 1908, Gabriel Lippmann, who made major contributions to color photography and three-dimensional displays, contemplated producing a display that provides a "window view upon reality" [Lippmann 1908]. Stephen Benton, one of the pioneers of holographic imaging, refined Lippmann's vision in the 1970s. He set out to design a scalable spatial display system with television-like characteristics, capable of delivering full color, 3D images with proper occlusion relationships. The display should provide images with binocular parallax (i.e., stereoscopic images) that can be viewed from any viewpoint without special glasses. Such displays are called *multiview autostereoscopic* since they naturally provide binocular and motion parallax for multiple observers. *3D video* usually refers to stored animated sequences, whereas *3D TV* includes real-time acquisition, coding, and transmission of dynamic scenes. In this paper we present the first end-to-end 3D TV system with 16 independent high-resolution views and autostereoscopic display.

Research towards the goal of end-to-end 3D TV started in Japan after the Tokyo Olympic Games in 1964 [Javidi and Okano 2002]. Most of that research focused on the development of binocular stereo cameras and stereo HDTV displays because the display of

multiple perspective views inherently requires a very high display resolution. For example, to achieve maximum HDTV output resolution with 16 distinct horizontal views requires $1920 \times 1080 \times 16$ or more than 33 million pixels, which is well beyond most current display technologies. It has only recently become feasible to deal with the high processing and bandwidth requirements of such high-resolution TV content.

In this paper we present a system for real-time acquisition, transmission, and high-resolution 3D display of dynamic multiview TV content. We use an array of hardware-synchronized cameras to capture multiple perspective views of the scene. We developed a fully distributed architecture with clusters of PCs on the sender and receiver side. We implemented several large, high-resolution 3D displays by using a multi-projector system and lenticular screens with horizontal parallax only. The system is scalable in the number of acquired, transmitted, and displayed video streams. The hardware is relatively inexpensive and consists mostly of commodity components that will further decrease in price. The system architecture is flexible enough to enable a broad range of research in 3D TV. Our system provides enough viewpoints and enough pixels per viewpoint to produce a believable and immersive 3D experience.

We make the following contributions:

Distributed architecture: In contrast to previous work in multiview video we use a fully distributed architecture for acquisition, compression, transmission, and image display.

Scalability: The system is completely scalable in the number of acquired, transmitted, and displayed views.

Multiview video rendering: A new algorithm efficiently renders novel views from multiple dynamic video streams on a cluster of PCs.

High-resolution 3D display: Our 3D display provides horizontal parallax with 16 independent perspective views at 1024×768 resolution.

Computational alignment for 3D displays: Image alignment and intensity adjustment of the 3D multiview display are completely automatic using a camera in the loop.

After an extensive discussion of previous work we give a detailed system overview, including a discussion of design choices and tradeoffs. Then we discuss the automatic system calibration using a camera in the loop. Finally, we present results, user experiences, and avenues for future work.

2 Previous Work and Background

The topic of 3D TV – with thousands of publications and patents – incorporates knowledge from multiple disciplines, such as image-based rendering, video coding, optics, stereoscopic displays, multi-projector displays, computer vision, virtual reality, and psychology. Some of the work may not be widely known across disciplines. There are some good overview books on 3D TV [Okoshi 1976; Javidi and Okano 2002]. In addition, we provide an extensive review of the previous work.

2.1 Model-Based Systems

One approach to 3D TV is to acquire multiview video from sparsely arranged cameras and to use some model of the scene for view interpolation. Typical scene models are per-pixel depth maps [Fehn et al. 2002; Zitnick et al. 2004], the visual hull [Matusik et al. 2000], or a prior model of the acquired objects, such as human body shapes [Carranza et al. 2003]. It has been shown that even coarse scene models improve the image quality during view synthesis [Gortler et al. 1996]. It is possible to achieve very high image quality with a two-layer image representation that includes automatically extracted boundary mattes near depth discontinuities [Zitnick et al. 2004].

One of the earliest and largest 3D video studios is the virtualized reality system of Kanade et al. [Kanade et al. 1997] with 51 cameras arranged in a geodesic dome. The Blue-C system at ETH Zürich consists of a room-sized environment with real-time capture and spatially-immersive display [Gross et al. 2003]. The Argus research project of the Air Force uses 64 cameras that are arranged in a large semi-circle [Javidi and Okano 2002, Chapter 9]. Many other, similar systems have been constructed.

All 3D video systems provide the ability to interactively control the viewpoint, a feature that has been termed *free-viewpoint video* by the MPEG Ad-Hoc Group on 3D Audio and Video (3DAV) [Smolic and Kimata 2003]. During rendering, the multiview video can be projected onto the model to generate more realistic view-dependent surface appearance [Matusik et al. 2000; Carranza et al. 2003]. Some systems also display low-resolution stereo-pair views of the scene in real-time.

Real-time acquisition of scene models for general, real-world scenes is very difficult and subject of ongoing research. Many systems do not provide real-time end-to-end performance, and if they do they are limited to simple scenes with only a handful of objects. We are using a dense lightfield representation that does not require a scene model, although we are able to benefit from it should it be available [Gortler et al. 1996; Buehler et al. 2001]. On the other hand, dense lightfields require more storage and transmission bandwidth. We demonstrate that these issues can be solved today.

2.2 Lightfield Systems

A lightfield represents radiance as a function of position and direction in regions of space free of occluders [Levoy and Hanrahan 1996]. The ultimate goal, which Gavin Miller called the "hyper display" [Miller 1995], is to capture a time-varying lightfield passing through a surface and emitting the same (directional) lightfield through another surface with minimal delay.

Early work in image-based graphics and 3D displays has dealt with static lightfields [Ives 1928; Levoy and Hanrahan 1996; Gortler et al. 1996]. In 1929, H. E. Ives proposed a photographic multi-camera recording method for large objects in conjunction with the first projection-based 3D display [Ives 1929]. His proposal bears some architectural similarities to our system, although modern technology allows us to achieve real-time performance.

Acquisition of dense, dynamic lightfields has only recently become feasible. Some systems use a bundle of optical fibers in front of a high-definition camera to capture multiple views simultaneously [Javidi and Okano 2002, Chapters 4 and 8]. The problem with single-camera systems is that the limited resolution of the camera greatly reduces the number and resolution of the acquired views.

Most systems – including ours – use a dense array of synchronized cameras to acquire high-resolution lightfields. The configuration and number of cameras is usually flexible. Typically, the cameras are connected to a cluster of PCs [Schirmacher et al. 2001; Naemura et al. 2002; Yang et al. 2002]. The Stanford multi-camera array [Wilburn et al. 2002] consists of up to 128 cameras and special-

purpose hardware to compress and store all the video data in real-time.

Most lightfield cameras allow interactive navigation and manipulation (such as "freeze frame" effects) of the dynamic scene. Some systems also acquire [Naemura et al. 2002] or compute [Schirmacher et al. 2001] per-pixel depth maps to improve the results of lightfield rendering. Our system uses 16 high-resolution cameras, real-time compression and transmission, and 3D display of the dynamic lightfield on a large multiview screen.

2.3 Multiview Video Compression and Transmission

Multiview video compression has mostly focused on static lightfields (e.g., [Magnor et al. 2003; Ramanathan et al. 2003]). There has been relatively little research on how to compress and transmit multiview video of dynamic scenes in real-time. A notable exception is the work by Yang et al. [2002]. They achieve real-time display from an 8×8 lightfield camera by transmitting only the rays that are necessary for view interpolation. However, it is impossible to anticipate all the viewpoints in a TV broadcast setting. We transmit all acquired video streams and use a similar strategy on the receiver side to route the videos to the appropriate projectors for display (see Section 3.3).

Most systems compress the multiview video off-line and focus on providing interactive decoding and display. An overview of some early off-line compression approaches can be found in [Javidi and Okano 2002, Chapter 8]. Motion compensation in the time domain is called *temporal encoding*, and disparity prediction between cameras is called *spatial encoding* [Tanimoto and Fuji 2003]. Zitnick et al. [Zitnick et al. 2004] show that a combination of temporal and spatial encoding leads to good results. The Blue-C system converts the multiview video into 3D "video fragments" that are then compressed and transmitted [Lamboray et al. 2004]. However, all current systems use a centralized processor for compression, which limits their scalability in the number of compressed views.

Another approach to multiview video compression, promoted by the European ATTEST project [Fehn et al. 2002], is to reduce the data to a single view with per-pixel depth map. This data can be compressed in real-time and broadcast as an MPEG-2 enhancement layer. On the receiver side, stereo or multiview images are generated using image-based rendering. However, it may be difficult to generate high-quality output because of occlusions or high disparity in the scene [Chen and Williams 1993]. Moreover, a single view cannot capture view-dependent appearance effects, such as reflections and specular highlights.

High-quality 3D TV broadcasting requires that all the views are transmitted to multiple users simultaneously. The MPEG 3DAV group [Smolic and Kimata 2003] is currently investigating compression approaches based on simultaneous temporal and spatial encoding. Our system uses temporal compression only and transmits all of the views as independent MPEG-2 video streams. We will discuss the tradeoffs in Section 3.2.

2.4 Multiview Autostereoscopic Displays

Holographic Displays It is widely acknowledged that the *hologram* was invented by Dennis Gabor in 1948 [Gabor 1948], although the French physicist Aimé Cotton first described holographic elements in 1901. Holographic techniques were first applied to image display by Leith and Upatnieks in 1962 [Leith and Upatnieks 1962]. In holographic reproduction, light from an illumination source is diffracted by interference fringes on the holographic surface to reconstruct the light wavefront of the original object. A hologram displays a continuous analog lightfield, and real-time acquisition and display of holograms has long been considered the "holy grail" of 3D TV.

Stephen Benton's Spatial Imaging Group at MIT has been pioneering the development of electronic holography. Their most recent device, the Mark-II Holographic Video Display, uses acousto-optic modulators, beamsplitters, moving mirrors, and lenses to create interactive holograms [St.-Hillaire et al. 1995]. In more recent systems, moving parts have been eliminated by replacing the acousto-optic modulators with LCD [Maeno et al. 1996], focused light arrays [Kajiki et al. 1996], optically-addressed spatial modulators [Stanley et al. 2000], or digital micromirror devices [Huebschman et al. 2003].

All current holo-video devices use single-color laser light. To reduce the amount of display data they provide only horizontal parallax. The display hardware is very large in relation to the size of the image (which is typically a few millimeters in each dimension). The acquisition of holograms still demands carefully controlled physical processes and cannot be done in real-time. At least for the foreseeable future it is unlikely that holographic systems will be able to acquire, transmit, and display dynamic, natural scenes on large displays.

Volumetric Displays Volumetric displays use a medium to fill or scan a three-dimensional space and individually address and illuminate small voxels [McKay et al. 2000; Favalora et al. 2001]. Actuality Systems (www.actuality-systems.com) and Neos Technologies (www.neostech.com) sell commercial systems for applications such as air-traffic control or scientific visualization. However, volumetric systems produce transparent images that do not provide a fully convincing three-dimensional experience. Furthermore, they cannot correctly reproduce the lightfield of a natural scene because of their limited color reproduction and lack of occlusions. The design of large-size volumetric displays also poses some difficult obstacles.

Akeley et al. [Akeley et al. 2004] developed an interesting fixed-viewpoint volumetric display that maintains view-dependent effects such as occlusion, specularity, and reflection. Their prototype uses beam-splitters to emit light at focal planes at different physical distances. Two such devices are needed for stereo viewing. Since the head and viewing positions remain fixed, this prototype is not a practical 3D display solution. However, it serves well as a platform for vision research.

Parallax Displays Parallax displays emit spatially varying directional light. Much of the early 3D display research focused on improvements to Wheatstone's stereoscope. In 1903, F. Ives used a plate with vertical slits as a barrier over an image with alternating strips of left-eye/right-eye images [Ives 1903]. The resulting device is called a *parallax stereogram*. To extend the limited viewing angle and restricted viewing position of stereograms, Kanolt [Kanolt 1918] and H. Ives [Ives 1928] used narrower slits and smaller pitch between the alternating image stripes. These multiview images are called *parallax panoramagrams*.

Stereograms and panoramagrams provide only horizontal parallax. In 1908, Lippmann proposed using an array of spherical lenses instead of slits [Lippmann 1908]. This is frequently called a "fly's-eye" lens sheet, and the resulting image is called an *integral photo-*

graph. An integral photograph is a true planar lightfield with directionally varying radiance per pixel (lenslet).

Integral lens sheets can be put on top of high-resolution LCDs [Nakajima et al. 2001]. Okano et al. [Javidi and Okano 2002, Chapter 4] connect an HDTV camera with fly's-eye lens to a high-resolution (1280×1024) LCD display. However, the resolution of their integral image is limited to 62×55 pixels. To achieve higher output resolution, Liao et al. [Liao et al. 2002] use a 3×3 projector array to produce a small display with 2872×2150 pixels. Their integral display with three views of horizontal and vertical parallax has a resolution of 240×180 pixels.

Integral photographs sacrifice significant spatial resolution in both dimensions to gain full parallax. Researchers in the 1930s introduced the *lenticular sheet*, a linear array of narrow cylindrical lenses called *lenticules*. This reduces the amount of image data by giving up vertical parallax. Lenticular images found widespread use for advertising, CD covers, and postcards [Okoshi 1976]. This has lead to improved manufacturing processes and the availability of large, high-quality, and very inexpensive lenticular sheets.

To improve the native resolution of the display, H. Ives invented the *multi-projector lenticular display* in 1931. He painted the back of a lenticular sheet with diffuse paint and used it as a projection surface for 39 slide projectors [Ives 1931]. Different arrangements of lenticular sheets and multi-projector arrays can be found in [Okoshi 1976, Chapter 5]. Based on this description we implemented both rear-projection and front-projection 3D display prototypes with a linear array of 16 projectors and lenticular screens (see Section 3.4). The high output resolution (1024×768), the large number of views (16), and the large physical dimension ($6' \times 4'$) of our display lead to a very immersive 3D experience.

Other research in parallax displays includes *time-multiplexed* (e.g., [Moore et al. 1996]) and *tracking-based* (e.g., [Perlin et al. 2000]) systems. In time-multiplexing, multiple views are projected at different time instances using a sliding window or LCD shutter. This inherently reduces the frame rate of the display and may lead to noticeable flickering. Head-tracking designs are mostly used to display stereo images, although it could also be used to introduce some vertical parallax in multiview lenticular displays.

Today's commercial autostereoscopic displays use variations of parallax barriers or lenticular sheets placed on top of LCD or plasma screens (www.stereo3d.com). Parallax barriers generally reduce some of the brightness and sharpness of the image. The highest resolution flat-panel screen available today is the IBM T221 LCD with about 9 million pixels. Our projector-based 3D display currently has a native resolution of 12 million pixels. We believe that new display media – such as organic LEDs or nanotube field-emission displays (FEDs) – will bring flat-panel multiview 3D displays within consumer reach in the foreseeable future.

2.5 Multi-Projector Displays

Scalable multi-projector display walls have recently become popular [Li et al. 2002; Raskar et al. 1998]. These systems offer very high resolution, flexibility, excellent cost-performance, scalability, and large-format images. Graphics rendering for multi-projector systems can be efficiently parallelized on clusters of PCs using, for example, the Chromium API [Humphreys et al. 2002]. Projectors also provide the necessary flexibility to adapt to non-planar display geometries [Raskar et al. 1999].

Precise manual alignment of the projector array is tedious and becomes downright impossible for more than a handful of projectors or non-planar screens. Some systems use cameras in the loop to automatically compute relative projector poses for automatic alignment [Raskar et al. 1999; Li et al. 2002]. Liao et al. [Liao et al. 2002] use a digital camera mounted on a linear 2-axis stage in their multi-projector integral display system. We use a static camera for automatic image alignment and brightness adjustments of the projectors (see Section 3.5).

3 System Architecture

Figure 2 shows a schematic representation of our 3D TV system. The *acquisition* stage consists of an array of hardware-synchronized cameras. Small clusters of cameras are connected to *producer* PCs. The producers capture live, uncompressed video streams and encode them using standard MPEG coding. The compressed video streams are then broadcast on separate channels over a *transmission* network, which could be digital cable, satellite TV, or the Internet. On the receiver side, individual video streams are decompressed by *decoders*. The decoders are connected by network (e.g., gigabit ethernet) to a cluster of *consumer* PCs. The consumers render the appropriate views and send them to a standard 2D, stereo-pair 3D, or multiview 3D display. In our current implementation, each consumer corresponds to a projector in the display and needs to project a slightly different viewpoint. A dedicated *controller* broadcasts the virtual view parameters to decoders and consumers. The controller is connected to a camera placed in the viewing area for automatic display calibration.

The system consists mostly of commodity components that are readily available today. The fully distributed processing makes it scalable in the number of acquired, transmitted, and displayed views. Note that the overall architecture of our system accommodates different display types (e.g., multiview 3D, head-mounted 2D stereo, or regular 2D).

3.1 Acquisition

Each camera captures progressive high-definition video in real-time. We are using 16 Basler A101fc color cameras (www.baslerweb.com) with 1300×1030, 8 bits per pixel CCD sensors. The cameras are connected by IEEE-1394 (FireWire) High Performance Serial Bus to the producer PCs. The maximum transmitted frame rate at full resolution is 12 frames per second. Two cameras each are connected to one of eight producer PCs. All PCs in our prototype have 3 GHz Pentium 4 processors, 2 GB of RAM, and run Windows XP.

We chose the Basler camera primarily because it has an external trigger that allows for complete control over the video timing. We have built a PCI card with custom programmable logic devices (CPLD) that generates the synchronization signal for all cameras. All 16 cameras are individually connected to the card, which is plugged into one of the producer PCs. Although it would be possible to use software synchronization [Yang et al. 2002], we consider precise hardware synchronization essential for dynamic scenes. Note that the price of the acquisition cameras can be high, since they will be mostly used in TV studios.

We arranged the 16 cameras in a regularly spaced linear array (see Figure 1 left). The optical axis of each camera is roughly perpendicular to a common camera plane. It is impossible to align multiple cameras precisely, so we use standard calibration procedures [Zhang 2000] to determine the intrinsic and extrinsic camera parameters.

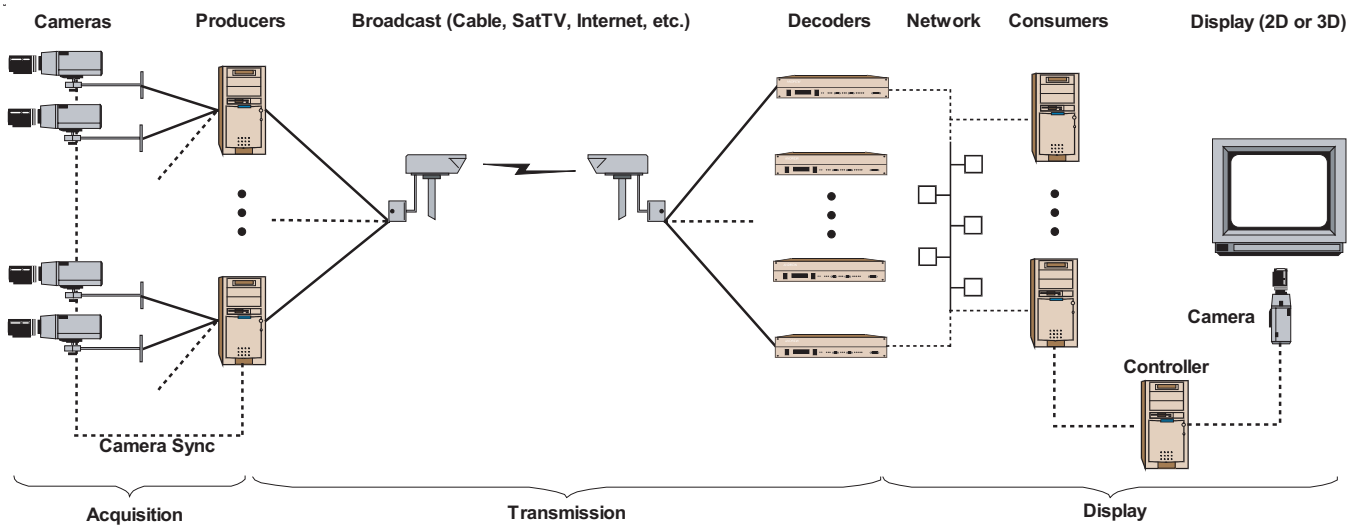

Figure 2: *A scalable end-to-end 3D TV system.*

In general, the cameras can be arranged arbitrarily because we are using lightfield rendering in the consumer to synthesize new views (see Section 3.3). A densely spaced array provides the best lightfield capture, but high-quality reconstruction filters could be used if the lightfield is undersampled [Stewart et al. 2003].

3.2 Transmission

Transmitting 16 uncompressed video streams with 1300×1030 resolution and 24 bits per pixel at 30 frames per second requires 14.4 Gb/sec bandwidth, which is well beyond current broadcast capabilities. For compression and transmission of dynamic multiview video data there are two basic design choices. Either the data from multiple cameras is compressed using spatial or spatio-temporal encoding, or each video stream is compressed individually using temporal encoding [1]. The first option offers higher compression, since there is a lot of coherence between the views. However, it requires that multiple video streams are compressed by a centralized processor. This compression-hub architecture is not scalable, since the addition of more views will eventually overwhelm the internal bandwidth of the encoder. Consequently, we decided to use temporal encoding of individual video streams on distributed processors.

This strategy has other advantages. Existing broadband protocols and compression standards do not need to be changed for immediate real-world 3D TV experiments and market studies. Our system can plug into today's digital TV broadcast infrastructure and co-exist in perfect harmony with 2D TV. Similar to HDTV, the introduction of 3D TV can proceed gradually, with one 3D channel at first and more to follow, depending on market demand. Note, however, that our transmission strategy is particular to broadcasting. Other applications (e.g., peer-to-peer 3D video conferencing) have different requirements, and we plan to investigate them in the future.

Another advantage of using existing 2D coding standards is that the codecs are well established and widely available. Tomorrow's digital TV set-top box could contain one or many decoders, depending whether the display is 2D or multiview 3D capable. Note that our system can adapt to other 3D TV compression algorithms [Fehn et al. 2002], as long as multiple views can be encoded (e.g,. into 2D video plus per-pixel depth maps [Flack et al. 2003]), transmitted, and decoded on the receiver side.

Because we did not have access to digital broadcast equipment, we implemented the modified architecture shown in Figure 3. Eight

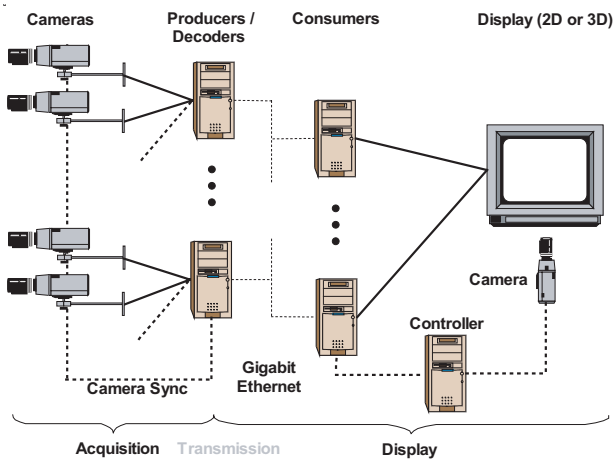

Figure 3: *Modified implementation for compression, transmission, and decoding in our prototype system.*

producer PCs are connected by gigabit ethernet to eight consumer PCs. Video streams at full camera resolution (1300×1030) are encoded with MPEG-2 and immediately decoded on the producer PCs. This essentially corresponds to a broadband network with infinite bandwidth and almost zero delay. We plan to introduce a more realistic broadband network simulation in the future. The gigabit ethernet provides all-to-all connectivity between decoders and consumers, which is important for our distributed rendering and display implementation.

3.3 Decoder and Consumer Processing

The receiver side is responsible for generating the appropriate images to be displayed. The system needs to be able to provide all possible views (i.e., the whole lightfield) to the end users at every time instance. The display controller requests one or more virtual

[1] Temporal encoding also uses spatial encoding within each frame, but not between views.

views by specifying the parameters of virtual cameras, such as position, orientation, field-of-view, and focal plane. In this discussion we assume that the user is not interactively navigating the lightfield and that the parameters of the virtual views remain fixed for a particular display.

The decoders receive a compressed video stream, decode it, and store the current uncompressed source frame in a buffer (see Figure 4). Each consumer has a *virtual video buffer (VVB)* with data

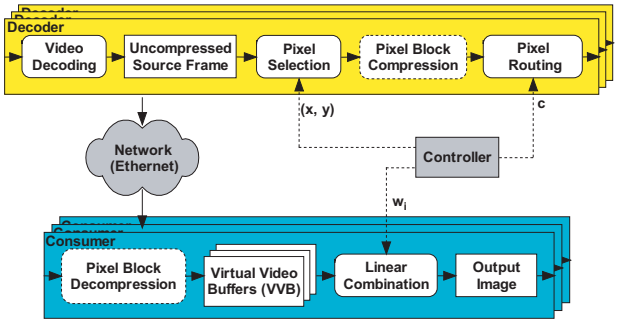

Figure 4: *Decoder and consumer processing.*

from *all* current source frames (i.e., all acquired views at a particular time instance). The consumer then generates a complete output image by processing image pixels from multiple frames in the VVB. Due to bandwidth and processing limitations it would be impossible for each consumer to receive the complete source frames from all the decoders. This would also limit the scalability of the system.

One possible implementation of our system uses a one-to-one mapping of cameras to projectors. In this case, the images need to be rectified using the camera calibration parameters, since the cameras are not accurately aligned. This approach is very simple and scales well, but the one-to-one mapping is not very flexible. For example, the cameras need to be equally spaced, which is hard to achieve in practice. Moreover, this method cannot handle the case when the number of cameras and projectors is not the same.

Another, more flexible approach is to use image-based rendering to synthesize views at the correct virtual camera positions. We are using unstructured lumigraph rendering [Buehler et al. 2001] on the consumer side. As in regular lightfield rendering, the geometric proxy for the scene is a single plane that can be set arbitrarily. We choose the plane that is roughly in the center of our depth of field. The virtual viewpoints for the projected images are chosen at even spacings.

Similar to [Yang et al. 2002], we observe that the contributions of the source frames to the output image of each consumer can be determined in advance. We now focus on the processing for one particular consumer, i.e., one particular view. For each pixel $o(u,v)$ in the output image, the display controller can determine the view number v and the position (x,y) of each source pixel $s(v,x,y)$ that contributes to it.

To generate output views from incoming video streams, each output pixel is a *linear combination* of k source pixels:

$$o(u,v) = \sum_{i=0}^{k} w_i s(v,x,y). \qquad (1)$$

The blending weights w_i can be pre-computed by the controller based on the virtual view information. The controller sends the positions (x,y) of the k source pixels to each decoder v for *pixel selection*. The index c of the requesting consumer is sent to the decoder for *pixel routing* from decoders to the consumer. Optionally, multiple pixels can be buffered in the decoder for *pixel block compression* before being sent over the network. The consumer decompresses the pixel blocks and stores each pixel in VVB number v at position (x,y).

Each output pixel requires pixels from k source frames. That means that the maximum bandwidth on the network to the VVB is k times the size of the output image times the number of frames per second (fps). For example, for $k = 3$, 30 fps and HDTV output resolution (1280×720 at 12 bits per pixel), the maximum bandwidth is 118 MB/sec. This can be substantially reduced if pixel block compression is used, at the expense of more processing. To provide scalability it is important that this bandwidth is independent of the total number of transmitted views, which is the case in our system. Note that we are using the term *pixel* loosely. It means typically one pixel, but it could also be an average of a small, rectangular block of pixels.

The processing requirements in the consumer are extremely simple. It needs to compute equation (1) for each output pixel. The weights are precomputed and stored in a lookup table. The memory requirements are k times the size of the output image. In our example above this corresponds to 4.3 MB. Assuming simple (lossless) pixel block compression, consumers can easily be implemented in hardware. That means that decoders, networks, and consumers could be combined on one printed circuit board or mass produced using an Application-Specific Integrated Circuit (ASIC). We may pursue this idea in the future.

So far we have assumed that the virtual views requested by the user / display are static. Note, however, that all the source views are sent over the broadcast network. The controller could update the lookup tables for pixel selection, routing, and blending dynamically. This would allow navigation of the lightfield similar to real-time lightfield cameras with random-access image sensors [Yang et al. 2002; Ooi et al. 2001], except that the frame buffers are on the receiver side. We plan to implement interactive navigation of the dynamic lightfield in the future.

3.4 3D Display

Figure 5 shows a diagram of our multi-projector 3D displays with lenticular sheets. We use 16 NEC LT-170 projectors with

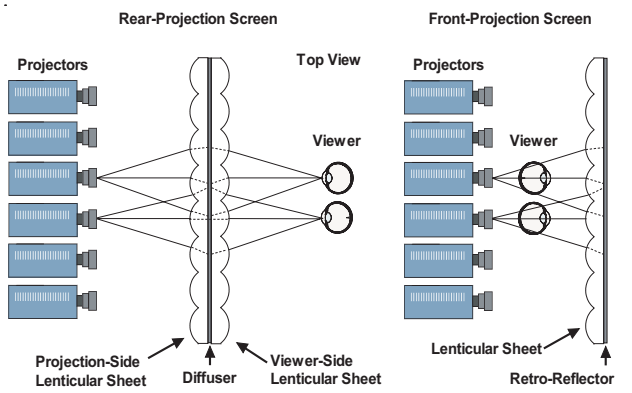

Figure 5: *Projection-type lenticular 3D displays.*

1024×768 native output resolution. Note that this is less than the resolution of our acquired and transmitted video, which we maintain at 1300×1030 pixels. However, HDTV projectors are still much more expensive than commodity projectors.

We chose this projector because of its compact form factor. Okoshi [1976] proposes values for optimal projector separation and lens pitch. Ideally, the separations between cameras and projectors are equal. We tried to match them approximately, which required mounting the projectors in three separate rows (see Figure 1 left). The offset in the vertical direction between neighboring projectors leads to a slight loss of vertical resolution in the final image.

We use eight consumer PCs and dedicate one of them as the controller. The consumers are identical to the producers except for a dual-output graphics card that is connected to two projectors. The graphics card is used only as an output device, since we perform the lightfield rendering in software.

For the rear-projection system (Figure 5 left), two lenticular sheets are mounted back-to-back with optical diffuser material in the center. We use a flexible rear-projection fabric from Da-Lite Screen Company (www.da-lite.com). The back-to-back lenticular sheets and the diffuser fabric were composited by Big3D Corp. (www.big3d.com) using transparent resin that was UV-hardened after hand-alignment. The front-projection system (Figure 5 right) uses only one lenticular sheet with a retro-reflective front-projection screen material from Da-Lite mounted on the back. Figure 1 shows photographs of both rear- and front-projection displays.

The projection-side lenticular sheet of the rear-projection display acts as a light multiplexer, focusing the projected light as thin vertical stripes onto the diffuser. A closeup of the lenticular sheet is shown in Figure 6. Considering each lenticule to be an ideal

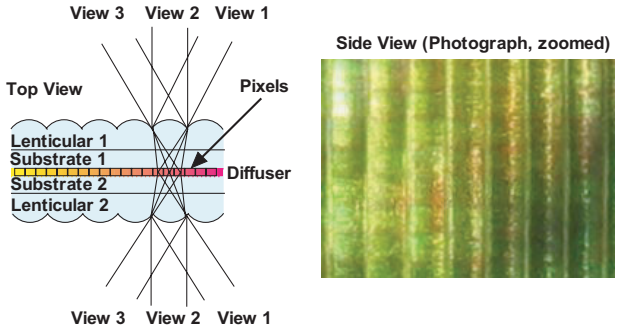

Figure 6: *Formation of vertical stripes on the diffuser of the rear-projection display. The closeup photograph (right) shows the lenticules and stripes from one viewpoint.*

pinhole camera, the stripes capture the view-dependent radiance of a three-dimensional lightfield (2D position and azimuth angle). The viewer-side lenticular sheet acts as a light de-multiplexer and projects the view-dependent radiance back to the viewer. Note that the single lenticular sheet of the front-projection screen both multiplexes and de-multiplexes the light.

The two key parameters of lenticular sheets are the field-of-view (FOV) and the number of lenticules per inch (LPI). We use $72'' \times 48''$ lenticular sheets from Microlens Technologies (www.microlens.com) with 30 degrees FOV and 15 LPI. The optical design of the lenticules is optimized for multiview 3D display. The number of *viewing zones* of a lenticular display are related to its FOV (see Figure 7). In our case, the FOV is 30 degrees, leading to $180/30 = 6$ viewing zones. At the border between two neighboring viewing zones there is an abrupt view-image change (or "jump") from view number 16 to view number one. The only remedy for this problem is to increase the FOV of the display. Note that each subpixel (or thin vertical stripe) in Figure 7 is projected from a different projector, and each projector displays images from a different view.

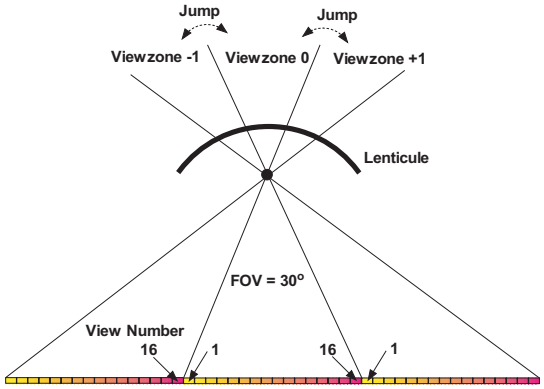

Figure 7: *The limited field-of-view of the lenticular sheets leads to sudden jumps of the image at the border of neighboring viewing zones.*

3.5 Display Calibration

As mentioned above, automatic projector calibration for the 3D display is very important. We first find the relationship between rays in space and pixels in the projected images by placing a camera on the projection side of the screen. Then we equalize the intensities of the projectors. For both processes, the display is covered with a diffuse screen material.

We use standard computer vision techniques [Raskar et al. 1999; Li et al. 2002] to find the mapping of points on the display to camera pixels, which (up to unknown scale) can be expressed by a 3×3 homography matrix. The largest common display area is computed by fitting the largest rectangle of a given aspect ratio (e.g., 4:3) into the intersection of all projected images.

Even for one projector, the intensities observed by the camera vary throughout the projected image. Moreover, different projectors may project images of vastly different intensities. Our calibration procedure works as follows. First, we project a white image in the common rectangle with each projector. We record the minimum intensity in this image for each projector and then we determine the minimum of those values across all projectors. This is the maximum intensity that we use for equalization. Next, we iteratively adjust the intensity of the image for each projector until the observed image has even maximum intensity. This is possible because we know the correspondence between the camera pixels and the pixels of each projector. This process yields image-intensity masks for each projector. It is only an approximate solution, since the response of the projectors for different intensities is generally nonlinear.

In the rear-projection system, a translation of the lenticular sheets with respect to each other leads to an apparent rotation of the viewing zones (see Figure 8). In order to estimate this horizontal shift

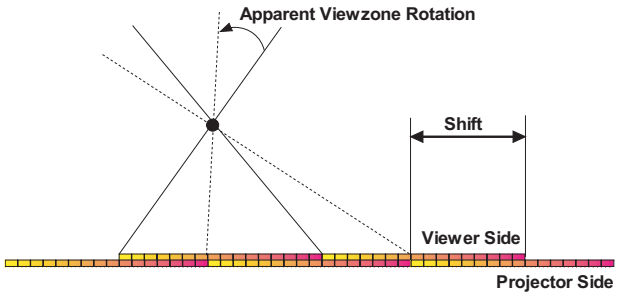

Figure 8: *Apparent viewing zone rotation for rear-projection due to a shift of the lenticular screens.*

we turn on each of the 16 projectors separately and measure the response on the viewer side of the display using the camera. The camera is placed approximately in the center of the display at the same distance to the screen as the projectors. We observe with which projector we achieve maximum brightness in the camera. We call this the apparent central projector. The image data is then re-routed between decoders (producers) and consumers such that the apparent central projector receives the images of the central camera.

4 Results

It is impossible to convey the impression of dynamic 3D TV on paper. Even the companion video – which is of course monocular – cannot do justice to the experience. Figure 9 shows four images that were taken at different positions on the viewer side of the display (top row) and the corresponding images of the camera array (bottom row). The parallax of the box in the foreground, the file cabinet on the right, and the door in the background are especially noticeable. Note that the color reproduction between the images of the displayed scene and the actual scene is quite similar.

Figures 10 and 11 show multiple viewpoints for a more dynamic scene. Note the increased parallax of objects (in this case the arms and legs) that are aimed straight at the camera array. Also note the relative change of positions between the painting on the back wall and the plant in Figure 11.

The frame rate of our end-to-end system is about 12 frames per second, which is the maximum frame rate of the cameras. However, the graphics cards and projectors are not synchronized, which leads to an increased motion blur for fast movements in the scene. In the future we plan to use high-end graphics cards with genlock capability.

All the images on the viewer side were taken from our front-projection display, which currently has the better image quality than the rear-projection display. This is mainly due to a slight rotation between the double-lenticular sheets, which leads to disturbing Moiré artifacts on the screen. The front-projection display does not have this problem. On the other hand, the front-projection system has more difficulty to represent pure blacks, and brightness or color variations between the projected images are more apparent. Note that the blur on the 3D display is quite prominent. This is due to the crosstalk between subpixels of different projectors and the light diffusion in the substrate of the lenticular sheets.

The feedback from early users of the system has been mostly positive. However, it is notable how the image quality problems of earlier prototype 3D displays distracted from the 3D experience. We believe the current image quality is acceptable, although it does not yet reach the quality of HDTV. Many of the remaining quality problems can be addressed in the future. For example, the lenticular sheet with 15 LPI shows some visible vertical lines. They would vanish if we were to increase the number of lenticules per inch.

We found that dynamic scenes – such as bouncing balls or jumps – are most fun to watch, especially in combination with the freeze-frame feature. Most users are surprised at the convincing 3D effect, and some of them keep coming back frequently. It should be noted, however, that this 3D display does not offer the kind of "in-your-face" 3D experience that one may find in 3D movie theaters with stereo glasses. Instead, it is really more like looking through a window at a scene in the distance.

5 Conclusions and Future Work

Most of the key ideas for the 3D TV system presented in this paper have been known for decades, such as lenticular screens, multi-projector 3D displays, and camera arrays for acquisition. The main advance over previous systems is to a large extent technological; such a system simply could not have been built until cameras and projectors were inexpensive enough and until computation was fast enough. We believe our system is the first to provide enough viewpoints and enough pixels per viewpoint to produce an immersive and convincing 3D experience (at least in the horizontal direction) without special glasses. It is also the first system that provides this experience in real-time for dynamic scenes.

There is still much that we can do to improve the quality of the 3D display. As noted before, the rear-projection system exhibits Moiré artifacts that can only be corrected by very precise vertical alignment of the lenticular sheets. We also noticed that the type of screen material (diffuse or retro-reflective) has a huge influence on the quality and sharpness of either rear- or front-projection screens. We are also experimenting with different lenticular sheets to improve the FOV and sharpness of the display. In the end, we believe that all of these issues can be resolved with sufficient amount of engineering effort.

Another area of future research is to improve the optical characteristics of the 3D display computationally. We call this concept the *computational display*. First, we plan to estimate the light transport matrix (LTM) of our view-dependent display by projecting patterns and observing them with a camera array on the viewer side. Knowing the LTM of the display will then allow us to modify the projected images to improve the quality. For example, we could change the displayed images for different areas in the viewing zone. The viewing-side cameras could also be replaced by a user who can tune the display parameters using a remote control to find the best viewing condition for the current viewpoint. The system could also try to optimize the display for as many users at different viewpoints as possible.

Another area of future research is precise color reproduction of natural scenes on multiview displays. Color reproduction standards for standard monitors have been issued by the International Color Consortium (www.color.org). However, very little work exists for color reproduction on multi-projector and multiview displays (the research by Stone [2001] is a notable exception). We plan to use the measured LTM and the viewer-side camera array to color-calibrate our multiview 3D display.

Another new and exciting area of research is high-dynamic range 3D TV. High-dynamic range cameras are being developed commercially and have been simulated using stereo cameras [Kang et al. 2003]. True high-dynamic range displays have also been developed [Seetzen et al. 2004]. We plan to extend these methods to multiview camera arrays and 3D displays.

In principle, our system could be used for tele-conferencing. The overall delay (from acquisition to display) is less than one second. However, we no longer could afford to transmit 16-channel multiview video over peer-to-peer connections. We plan to investigate new multiview video compression techniques in the future. The broadcast or peer-to-peer networks will of course introduce issues of quality of service, bandwidth limitations, delays, buffering of multiple streams, and so on. We believe these issues are well understood in the broadcast community, and we plan to address them in the future.

We also plan to experiment with multiview displays for deformable display media, such as organic LEDs. If we know the orientation

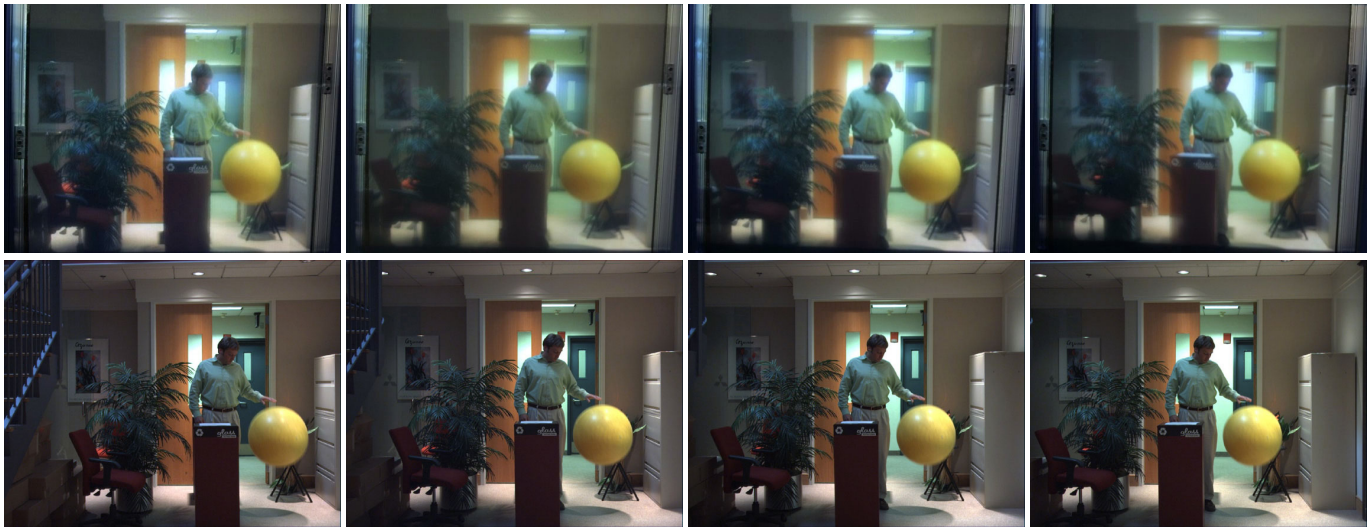

Figure 9: *Images of a scene from the viewer side of the display (top row) and as seen from some of the cameras (bottom row).*

Figure 10: *A more dynamic scene as displayed on the viewer side (top row) and as seen by the cameras (bottom row).*

Figure 11: *Another dynamic scene as seen from different viewpoints on the 3D display.*

and relative position of each display element we can render new virtual views by dynamically routing image information from the decoders to the consumers. As noted by Gavin Miller [Miller 1995], this would allow the design of "invisibility cloaks" by displaying view-dependent images on an object that would be seen if the object were not present. One could dynamically update these views using tiny cameras that are positioned around the object. Multiview cameras and displays that dynamically change their parameters, such as position, orientation, focus, or aperture, pose new and exciting research problems for the future.

6 Acknowledgments

We thank Joe Marks for valuable discussions and for his support throughout this project. We also thank Marc Levoy and the students in the Stanford computer graphics lab for stimulating discussions and useful suggestions. Marc Levoy pointed out some valuable references that were added to the paper. Also thanks to Leonard McMillan for initial discussions and ideas and to Jennifer Roderick Pfister for proofreading the paper.

References

AKELEY, K., WATT, S., GIRSHICK, A., AND BANKS, M. 2004. A stereo display prototype with multiple focal distances. *To appear in ACM Transaction on Graphics* (Aug.).

BUEHLER, C., BOSSE, M., MCMILLAN, L., GORTLER, S., AND COHEN, M. 2001. Unstructured lumigraph rendering. In *Computer Graphics*, SIGGRAPH 2001 Proceedings, 425–432.

CARRANZA, J., THEOBALT, C., MAGNOR, M., AND SEIDEL, H. 2003. Free-viewpoint video of human actors. *ACM Transactions on Graphics 22*, 3, 569–577.

CHEN, S. E., AND WILLIAMS, L. 1993. View interpolation for image synthesis. In *Computer Graphics*, SIGGRAPH 93 Proceedings, 279–288.

FAVALORA, G., DORVAL, R., HALL, D., M., M. G., AND NAPOLI, J. 2001. Volumetric three-dimensional display system with rasterization hardware. In *Stereoscopic Displays and Virtual Reality Systems VIII*, vol. 4297 of *SPIE Proceedings*, 227–235.

FEHN, C., KAUFF, P., DE BEECK, M. O., ERNST, F., IJSSELSTEIJN, W., POLLEFEYS, M., GOOL, L. V., OFEK, E., AND SEXTON, I. 2002. An evolutionary and optimised approach on 3D-TV. In *Proceedings of International Broadcast Conference*, 357–365.

FLACK, J., HARMAN, P., AND FOX, S. 2003. Low bandwidth stereoscopic image encoding and transmission. In *Stereoscopic Displays and Virtual Reality Systems X*, vol. 5006 of *Proceedings of SPIE*, 206–215.

GABOR, D. 1948. A new microscopic principle. *Nature*, 161 (May), 777–779.

GORTLER, S., GRZESZCZUK, R., SZELISKI, R., AND COHEN, M. 1996. The lumigraph. In *Computer Graphics*, SIGGRAPH 96 Proceedings, 43–54.

GROSS, M., WUERMLIN, S., NAEF, M., LAMBORAY, E., SPAGNO, C., KUNZ, A., KOLLER-MEIER, E., SVOBODA, T., GOOL, L. V., LANG, S., STREHLKE, K., MOERE, A. V., AND STAADT, O. 2003. blue-c: A spatially immersive display and 3D video portal for telepresence. *ACM Transactions on Graphics 22*, 3, 819–828.

HUEBSCHMAN, M., MUNJULURI, B., AND GARNER, H. R. 2003. Dynamic holographic 3-D image projection. *Optics Express*, 11, 437–445.

HUMPHREYS, G., HOUSTON, M., NG, Y., FRANK, R., AHERN, S., KIRCHNER, P., AND KLOSOWSKI, J. 2002. Chromium: A stream processing framework for interactive graphics on clusters. *ACM Transactions on Graphics 21*, 3, 693–703.

IVES, F. E., 1903. Parallax stereogram and process for making same. U.S. Patent Number 725,567, filed 1902, Apr.

IVES, H. E. 1928. A camera for making parallax panoramagrams. *Journal of the Optical Society of America*, 17 (Dec.), 435–439.

IVES, H. E. 1929. Motion pictures in relief. *Journal of the Optical Society of America*, 18 (Feb.), 118–122.

IVES, H. E. 1931. The projection of parallax panoramagrams. *Journal of the Optical Society of America*, 21 (July), 397–409.

JAVIDI, B., AND OKANO, F., Eds. 2002. *Three-Dimensional Television, Video, and Display Technologies*. Springer-Verlag.

KAJIKI, Y., YOSHIKAWA, H., AND HONDA, T. 1996. Three-dimensional display with focused light array. In *Practical Holography X*, vol. 2652 of *SPIE Proceedings*, 106–116.

KANADE, T., RANDER, P., AND NARAYANAN, P. 1997. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multimedia, Immersive Telepresence 4*, 1 (Jan.), 34–47.

KANG, S. B., UYTTENDAELE, M., WINDER, S., AND SZELISKI, R. 2003. High dynamic range video. *ACM Transactions on Graphics 22*, 3, 319–325.

KANOLT, C. W., 1918. Photographic method and apparatus. U.S. Patent Number 1,260,682, filed 1915, Mar.

LAMBORAY, E., WÜRMLIN, S., AND GROSS, M. 2004. Real-time streaming of point-based 3D video. In *To appear in: Proceedings of IEEE Virtual Reality*.

LEITH, E., AND UPATNIEKS, J. 1962. Reconstructed wavefronts and communication theory. *Journal of the Optical Society of America 52*, 10 (Oct.), 1123–1130.

LEVOY, M., AND HANRAHAN, P. 1996. Light field rendering. In *Computer Graphics*, SIGGRAPH 96 Proceedings, 31–42.

LI, K., CHEN, H., CHEN, Y., CLARK, D., COOK, P., DAMIANAKIS, S., ESSL, G., FINKELSTEIN, A., FUNKHOUSER, T., HOUSEL, T., KLEIN, A., LIU, Z., PRAUN, E., SAMANTA, R., SHEDD, B., SINGH, J. P., TZANETAKIS, G., AND ZHENG, J. 2002. Building and using a scalable display wall system. *IEEE Computer Graphics and Applications 20*, 4 (Dec.), 29–37.

LIAO, H., IWAHARA, M., HATA, N., SAKUMA, I., DOHI, T., KOIKE, T., MOMOI, Y., MINAKAWA, T., YAMASAKI, M., TAJIMA, F., AND TAKEDA, H. 2002. High-resolution integral videography autostereoscopic display using multi-projector. In *Proceedings of the Ninth International Display Workshop*, 1229–1232.

LIPPMANN, G. 1908. Epreuves reversibles donnant la sensation du relief. *Journal of Physics 7*, 4 (Nov.), 821–825.

MAENO, K., FUKAYA, N., NISHIKAWA, O., SATO, K., AND HONDA, T. 1996. Electroholographic display using 15-megapixel LCD. In *Practical Holography X*, vol. 2652 of *SPIE Proceedings*, 15–23.

MAGNOR, M., RAMANATHAN, P., AND GIROD, B. 2003. Multi-view coding for image-based rendering using 3-D scene geometry. *IEEE Trans. Circuits and Systems for Video Technology 13*, 11 (Nov.), 1092–1106.

MATUSIK, W., BUEHLER, C., RASKAR, R., GORTLER, S., AND MCMILLAN, L. 2000. Image-based visual hulls. In *Computer Graphics*, SIGGRAPH 2000 Proceedings, 369–374.

MCKAY, S., MAIR, G., MASON, S., AND REVIE, K. 2000. Membrane-mirrorbased autostereoscopic display for teleoperation and telepresence applications. In *Stereoscopic Displays and Virtual Reality Systems VII*, vol. 3957 of *SPIE Proceedings*, 198–207.

MILLER, G. 1995. Volumetric hyper-reality, a computer graphics holy grail for the 21st century? In *Proceedings of Graphics Interface '95*, Canadian Information Processing Society, 56–64.

MOORE, J. R., DODGSON, N., TRAVIS, A., AND LANG, S. 1996. Time-multiplexed color autostereoscopic display. In *Symposium on Steroscopic Displays and Applications VII*, vol. 2653 of *Proceedings of SPIE*, 1–9.

NAEMURA, T., TAGO, J., AND HARASHIMA, H. 2002. Real-time video-based modeling and rendering of 3D scenes. *IEEE Computer Graphics and Applications* (Mar.), 66–73.

NAKAJIMA, S., NAKAMURA, K., MASAMUNE, K., SAKUMA, I., AND DOHI, T. 2001. Three-dimensional medical imaging display with computer-generated integral photography. *Computerized Medical Imaging and Graphics 25*, 3, 235–241.

OKOSHI, T. 1976. *Three-Dimensional Imaging Techniques*. Academic Press.

OOI, R., HAMAMOTO, T., NAEMURA, T., AND AIZAWA, K. 2001. Pixel independent random access image sensor for real time image-based rendering system. In *IEEE International Conference on Image Processing*, vol. II, 193–196.

PERLIN, K., PAXIA, S., AND KOLLIN, J. 2000. An autostereoscopic display. In *SIGGRAPH 2000 Conference Proceedings*, vol. 33, 319–326.

RAMANATHAN, P., KALMAN, M., AND GIROD, B. 2003. Rate-distortion optimized streaming of compressed light fields. In *IEEE International Conference on Image Processing*, 277–280.

RASKAR, R., WELCH, G., CUTTS, M., LAKE, A., STESIN, L., AND FUCHS, H. 1998. The office of the future: A unified approach to image-based modeling and spatially immersive displays. In *Proceedings of SIGGRAPH 98*, 179–188.

RASKAR, R., BROWN, M., YANG, R., CHEN, W., WELCH, G., TOWLES, H., SEALES, B., AND FUCHS, H. 1999. Multi-projector displays using camera-based registration. In *IEEE Visualization*, 161–168.

SCHIRMACHER, H., MING, L., AND SEIDEL, H.-P. 2001. On-the-fly processing of generalized lumigraphs. In *Proceedings of Eurographics 2001*, vol. 20 of *Computer Graphics Forum*, Eurographics Association, 165–173.

SEETZEN, H., HEIDRICH, W., STUERZLINGER, W., WARD, G., WHITEHEAD, L., TRENTACOSTE, M., GHOSH, A., AND VOROZCOVS, A. 2004. High dynamic range display systems. *To appear in ACM Transaction on Graphics* (Aug.).

SMOLIC, A., AND KIMATA, H., 2003. Report on 3DAV exploration. ISO/IEC JTC1/SC29/WG11 Document N5878, July.

ST.-HILLAIRE, P., LUCENTE, M., SUTTER, J., PAPPU, R., C.J.SPARRELL, AND BENTON, S. 1995. Scaling up the MIT holographic video system. In *Proceedings of the Fifth International Symposium on Display Holography*, SPIE, 374–380.

STANLEY, M., CONWAY, P., COOMBER, S., JONES, J., SCATTERGOOD, D., SLINGER, C., BANNISTER, B., BROWN, C., CROSSLAND, W., AND TRAVIS, A. 2000. A novel electro-optic modulator system for the production of dynamic images from giga-pixel computer generated holograms. In *Practical Holography XIV and Holographic Materials VI*, vol. 3956 of *SPIE Proceedings*, 13–22.

STEWART, J., YU, J., GORTLER, S., AND MCMILLAN, L. 2003. A new reconstruction filter for undersampled light fields. In *Eurographics Symposium on Rendering*, ACM International Conference Proceeding Series, 150–156.

STONE, M. 2001. Color and brightness appearance issues in tiled displays. *Computer Graphics and Applications 21*, 6 (Sept.), 58–67.

TANIMOTO, M., AND FUJI, T., 2003. Ray-space coding using temporal and spatial predictions. ISO/IEC JTC1/SC29/WG11 Document M10410, Dec.

WILBURN, B., SMULSKI, M., LEE, H. K., AND HOROWITZ, M. 2002. The light field video camera. In *Media Processors 2002*, vol. 4674 of *SPIE*, 29–36.

YANG, J. C., EVERETT, M., BUEHLER, C., AND MCMILLAN, L. 2002. A real-time distributed light field camera. In *Proceedings of the 13th Eurographics Workshop on Rendering*, Eurographics Association, 77–86.

ZHANG, Z. 2000. A flexible new technique for camera calibration. *IEEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 11, 1330–1334.

ZITNICK, L., KANG, S. B., UYTTENDAELE, M., WINDER, S., AND SZELISKI, R. 2004. High-quality video view interpolation using a layered representation. *To appear in ACM Transaction on Graphics* (Aug.).