Virtual Studios: An Overview

Simon Gibbs, Costas Arapis, Christian Breiteneder, Vali Lalioti, Sina Mostafawy, and Josef Speier
GMD Institute for Media Communication

Virtual studio systems began as experimental prototypes that extended traditional chromakeying. Now commercial products based on graphics supercomputers are commonly used for broadcast production. We discuss this evolution and consider extensions, alternative approaches, and issues facing broadcasters who introduce virtual studio systems.

In video production, virtual studios provide seamless real-time compositing of live video with synthetic or natural imagery. This new technique, sometimes called third-person virtual reality, lets those observing the "mixed signal" see people and other physical objects combined with a virtual environment. The hardware and software needed to achieve this comprise a virtual studio system, while the imagery combined with the live video constitutes a virtual set.

Figure 1 shows 3DK, a typical virtual studio system developed at GMD (German National Research Center for Information Technology). A set of cameras equipped with tracking systems provide information about camera movement. The cameras produce foreground (FG) video signals, while an image generation system such as a graphics supercomputer produces corresponding background (BG) signals and, optionally, mask signals (Figure 2). The foreground and background are then composited, often by chromakeying, and the composited outputs (foreground+background) made available to downstream components such as mixers, special effects devices, monitors, and recorders. In addition, each composited signal is fed back to the corresponding camera's viewfinder to assist the camera operator in shot placement.

As Figure 1 shows, virtual studio systems divide naturally into three main subsystems: tracking, rendering, and compositing. For real-time performance, each uses special hardware capable of video-rate operation (50-60 Hz).

Background

Virtual studios use several video production techniques. We'll review them first and give more detailed technical descriptions later.

Chromakeying

Chromakeying, a staple of video production, provides a good starting point for understanding the historical development of virtual studios. In traditional chromakeying, the subject is shot against a constant background such as a blue curtain or screen. This "blue screen" shot then passes through a chromakeyer, where it is combined with a second shot containing the new background. Conceptually, chromakeyer operation is simple: Replace the foreground with the background in those places where the foreground contains a particular color known as the key color. (Strictly speaking, a key is associated with a region in color space rather than a single color.) The chromakeyer operator selects the key color or the chromakeyer itself may do so automatically.

Unfortunately, simple comparison of foreground pixels to the key color produces unnaturally hard edges around foreground objects (such as hair with a sculpted appearance). Also, shadows of foreground objects falling on the blue screen will be clipped out. Modern chromakey systems therefore apply techniques to preserve foreground detail and suppress artifacts such as blue fringes around foreground objects.

Despite chromakey systems' sophistication, their operation imposes a fundamental constraint: The foreground camera (and background camera, if the background is a camera shot) cannot move—it must be "locked off" for the shot's duration. Consider a typical chromakeying situation, the weather announcer and map. If the camera shooting the announcer pans, the announcer appears to slide over the (fixed) background; if the camera moves forward or back, or zooms, the announcer appears to grow or shrink. In other words, the spatial relationships existing between the two layers (such as a foreground object appearing left of a background object) are not consistently maintained. Fixing camera position helps maintain the coherence of foreground and background layers but is extremely limiting. As a result, a number of techniques have been developed to extend chromakeying's usefulness.

Synthevision

In the late 1980s, researchers at the Japanese broadcast company NHK developed a form of

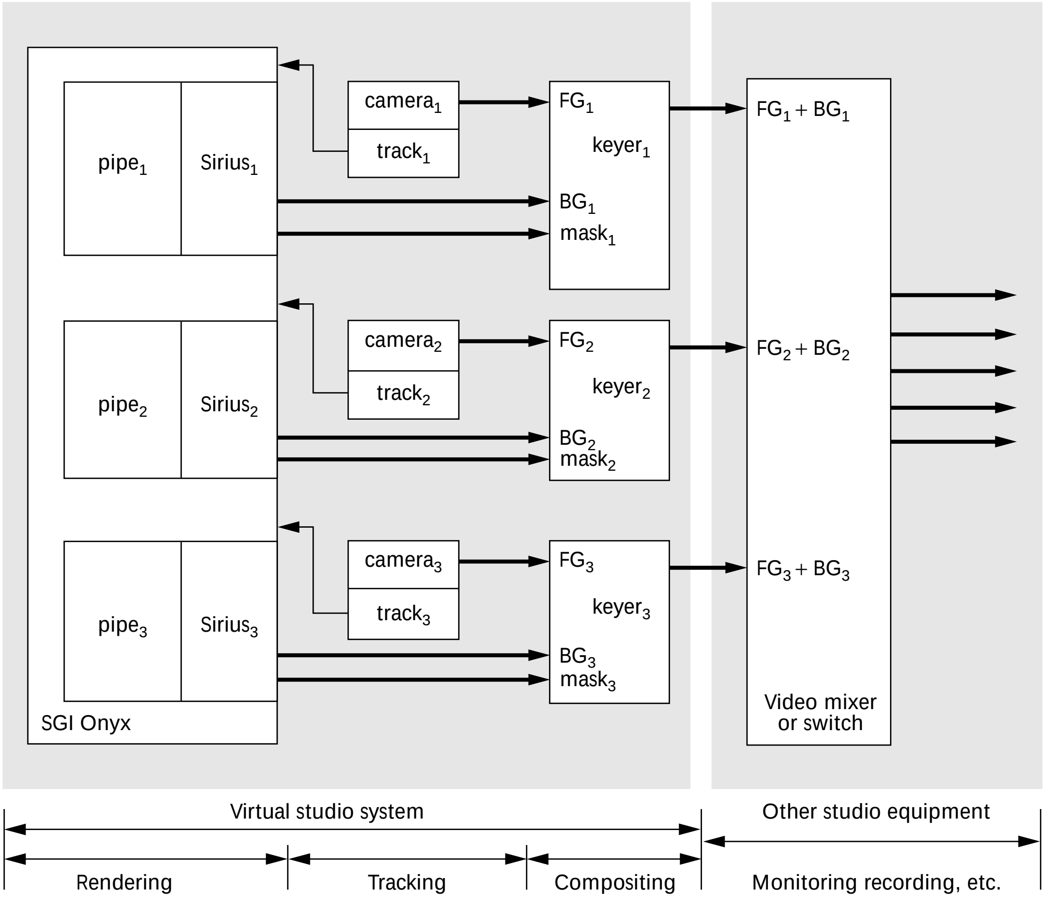


Figure 1. A virtual studio system based on the SGI Onyx and Sirius video board.

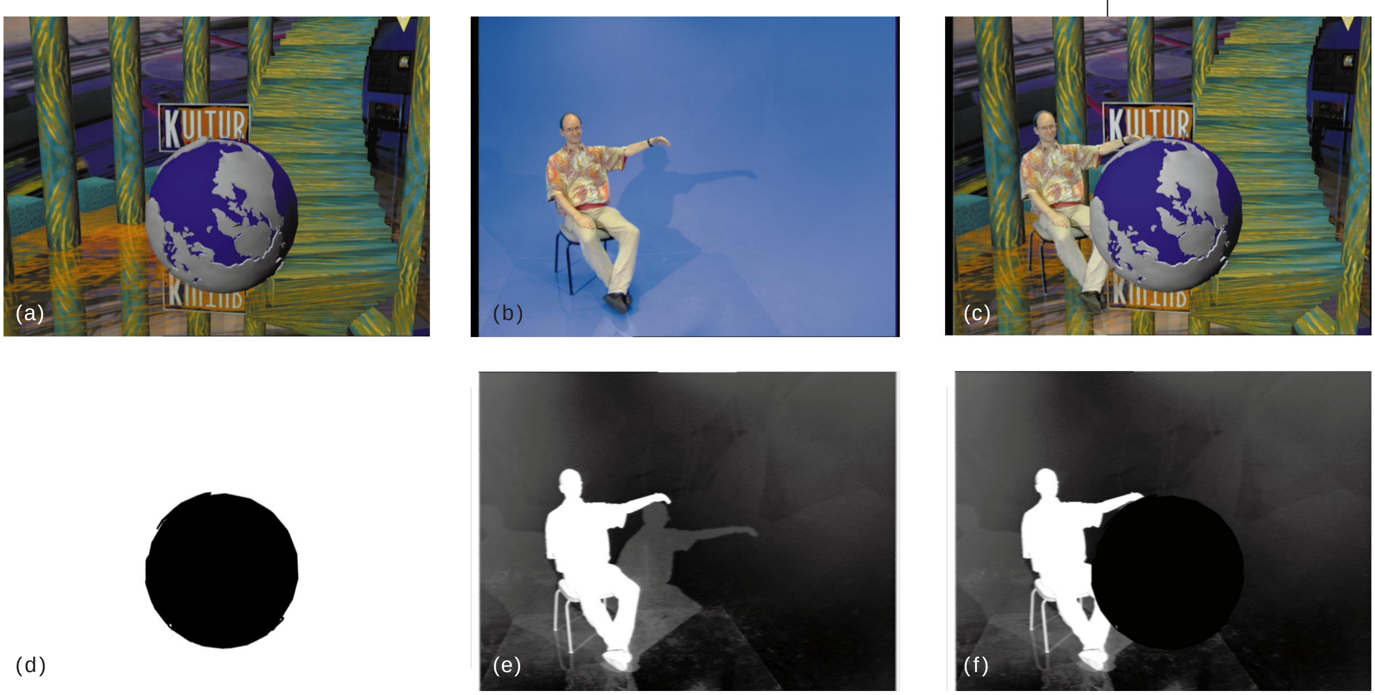


Figure 2. Video signals used in a virtual studio system: (a) background, (b) foreground, (c) background plus foreground, (d) foreground mask, (e) background mask, and (f) foreground mask plus background mask.

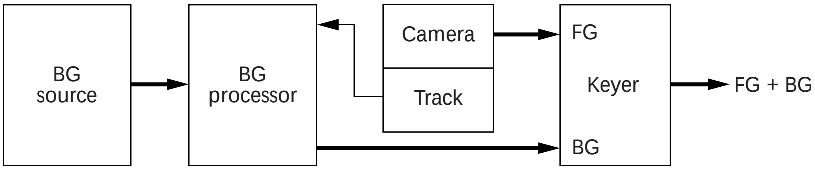


Figure 3. NHK's Synthevision.

chromakeying called Synthevision[1] (now a product). Sensors attached to the foreground camera produce pan, tilt, zoom, and focus data. The camera motions are simulated on the background layer in real time, which involves altering the background perspective to match that of the foreground. The foreground and modified background are then composited using a conventional chromakeyer (see Figure 3). The background consists of an HDTV (high-definition television, or Hi-Vision) still frame, and the larger size gives more image area to simulate panning, tilting, and zooming.

Synthevision first appeared during the Seoul Olympics in 1988, and NHK uses it daily for newscasts. Although the background images are generally still frames, the designers anticipated the true virtual studio and noted that "by using CG images in backgrounds, studio sets that cannot possibly exist can be used easily for program production."[1] NHK continues to experiment with camera tracking and extensions to chromakeying. For example, they use multiple video layers to provide realistic motion parallax and have incorporated real-time computer-generated sets in several productions.

Camera ganging

An alternative to coupling a real and a virtual camera is "ganging" two real cameras in a master-slave relationship (see Figure 4). Typically, the foreground camera operates as master, and the background camera duplicates any motion it makes. This technique is common in special-effects production and proves particularly useful if more than one layer contains live (as opposed to computer-generated) video. Possible forms of ganging include

- recording one camera's moves and using the data to drive a second camera,
- physically interconnecting two cameras so that the slave camera duplicates the master camera's motion as the operator moves, and
- controlling cameras by an external source (such as a computer) that directs them either simultaneously or at different times through a sequence of moves.

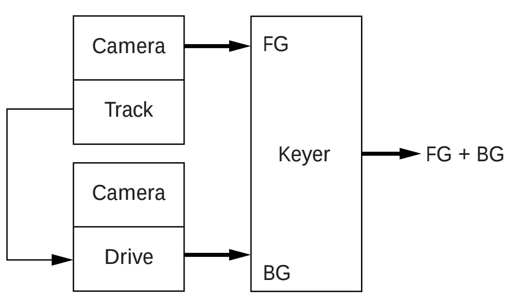


Figure 4. Camera ganging.

Prerendered virtual sets

If camera motion and settings are known, it is possible to prerender the virtual set and run the rendered material and the camera material through a chromakeyer. BBC has used this technique in newscasts,[2] rendering the virtual set field by field to a hard disk recorder and then mixing it with a live camera (showing the announcer) just prior to broadcast. This technique has two basic variants:

- Record the foreground and either track the camera or calculate camera motion via frame-by-frame analysis (not necessarily in real time). Then use the motion data to position the virtual camera and render the virtual set, then composite the prerecorded foreground and prerendered background.
- Specify camera motion, for example using animation software, and prerender the background and store it on disk or tape. Produce the foreground using a servo-control system that drives the foreground camera through the predetermined move sequence. Playing back the background in sync with the foreground permits live compositing (see Figure 5).

Early virtual studio systems

Real-time virtual sets were most likely first used in Japan in 1991. NHK used a prototype virtual studio system to produce "Nanospace", a science documentary. The show was broadcast in 1992 and, in hindsight, was ahead of its time. (For more about this work, see "An Image Compositing System Based on Virtual Cameras" in this issue, p.36.)

NHK's pioneering work contained the main elements of a virtual studio system: real-time background rendering (using an SGI VTX) coupled with real-time foreground tracking (using a sensor system developed in-house). Unfortunately, NHK's efforts were hampered by the graphics hardware performance then available. Commercial virtual studio systems did not arrive until after SGI's 1993 introduction of the RealityEngine2.

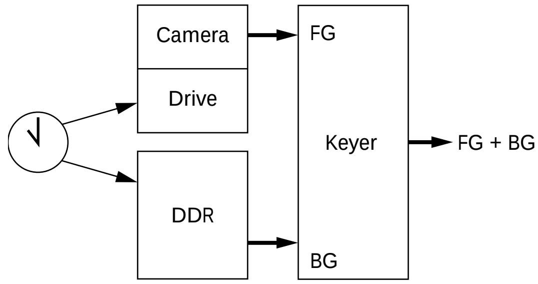


Figure 5. Prerendered virtual set. A frame sequence is stored on a digital disk recorder (DDR) and the corresponding sequence of camera motion parameters is stored in the camera drive unit. The two sequences are played back in sync to a reference time code signal.

Two other early virtual studio systems were developed in Europe: Platform by IMP, a company with expertise in real-time graphics for video production; and Electronic Set (ELSET), developed by the Mona Lisa project,[3,4] a European Union RACE project led by the Hamburg production company VAP. In 1995, Discreet Logic acquired Platform and now markets it as Vapour. ELSET is available from Accom.

Table 1, adapted from Hughes,[5] summarizes virtual studio development (our apologies for any errors or omissions). We've added the first public showing of our own virtual studio system, 3DK.

Table 1 clearly indicates that virtual studios systems are proliferating. Broadcasters show strong interest, and many are planning or already engaged in test productions. They are also starting to realize that virtual studios form just part of a much larger trend. Broadcasting industry issues range from the move from analog to digital technology, the future of HDTV, and the role of compression and digital transmission systems to interactivity and the Internet. Studios and production processes will likely undergo major restructuring over the next decade, and the adoption of virtual studio techniques constitutes just one element of this industry transformation.

Digital video technology

Virtual studio systems generally use the digital video standard CCIR 601, which specifies an

Table 1. Virtual studio chronology.

Date (Event)	Company (Product)
1988 (Seoul Olympics)	NHK (Synthevision—2D background)
1991	NHK (in-house system)
1992	Ultimatte (prerendered virtual set)
1993	BBC (in-house system—prerendered virtual set)
March 1994 (National Association of Broadcasters—NAB)	IMP (Platform)
September 1994 (International Broadcasting Convention—IBC)	IMP (Platform), VAP (ELSET)
November 1994	GMD (3DK)
April 1995 (NAB)	IMP (Platform), Accom (ELSET), RT-Set Ltd. (Larus, Otus) ElectroGIG (Reality Tracking) Softimage/INA (Virtual Theater/Hybrid Vision)
September 1995 (IBC)	IMP (Platform), Accom (ELSET), RT-Set Ltd. (Larus, Otus) ElectroGIG (Reality Tracking) Softimage (Virtual Theater) Orad (Virtual Set)
April 1996 (NAB)	Discreet Logic (Vapour), Accom (ELSET), RT-Set Ltd. / Chyron (Larus, Otus) ElectroGIG (Reality Tracking) Orad (Cyberset), Evans and Sutherland (MindSet) Vinten(VideoScape), Radamec (Virtual Scenario)
1997	Primarily the above with PC-based versions now appearing

uncompressed digital representation for analog YUV video. It uses a base sampling frequency of approximately 3.375 MHz and samples the analog components at various multiples of this frequency. For instance, "4:2:2" indicates that *Y* (the luminance component) is sampled at four times the base frequency while *UV* (the two chroma components) are sampled at two times the base frequency. Each component's sample size is 10 bits (older equipment uses 8 bits), yielding a data rate of 270 Mbps for 4:2:2. Frame sizes are 720×576 for PAL (Phase Alternate Line, a primarily European television standard) and 720×486 for NTSC (National Television System Committee); unfortunately, the pixel aspect ratio is not 1:1.

An extension to CCIR 601 specifies an interconnect standard (often called serial digital interface, or SDI) that uses normal coaxial cable to carry digital video signals. This industry-supported interface allows the connection of a wide variety of video equipment including digital video tape recorders (D1, D5, and digital Betacam), disk recorders, mixers, routing switches, frame synchronizers, and converters. The MPEG-2 encoders found in digital broadcast systems often have serial digital interfaces.

The 4:2:2 format, and 4:0:0 for luma-only signals such as keys, is a mainstay of video postproduction environments. The D1 and D5 video tape recorders and the new disk-based storage systems permit building up complex multilayer composites without the generation loss found in analog systems. This format also works well for virtual studio systems, since background generation is typically a digital process. In other words, the foreground, background, and mask signals used in a virtual studio system can all be based on the same digital video format. No conversions between analog and digital are required (other than initial digitization of the camera signal). Of course, after production and any editing or subsequent processing, the signal may be converted to analog for distribution.

Blue room design and lighting

Since virtual studio systems allow camera movement, the design of the blue backing becomes more important than with "locked off" cameras. For a weather announcer and map, one can probably get away with a fairly small blue screen (perhaps 2×1.5 meters) placed behind the announcer. Once the camera is free to move, a larger blue screen gives better results. Furthermore, long shots require blue not only behind the announcer but also on the floor. For this reason, shooting with a virtual studio system often occurs in a blue room (or "blue box") consisting of one or more blue walls and a blue floor.

Lighting is probably the most important consideration in blue room design. Good compositing depends on uniform background lighting. Adding a floor complicates lighting—because the floor and walls have different orientations to light sources, they appear as slightly different shades of blue. This is particularly noticeable along the edges and may show through as shadowy bands in the composite. Curved sections between walls and the floor and between the walls soften these edges, and using the same material for floor and walls makes lighting easier. Although adding curved sections diminishes the blue room's usable area, the virtual set can be much larger than the physical set.

The physical set need not be blue—chromakeyers and other compositing devices work with red or green levels, luminance levels, or even arbitrary colors. However, blue offers several advantages. Blue complements flesh tones and so better preserves them (at least by some compositing processes); also, a slight blue tinge around objects is less objectionable than a green tinge. Actors probably find working in a blue room more pleasant than working in a strongly saturated green or red room.

In addition to the need for overall uniform lighting on the backing, other factors influence blue room lighting. First, a lot of blue light reflects onto objects. This "blue spill" can be removed during the compositing process but results in a loss of side and back lighting on the subject that should be compensated for with additional lighting. Second, the intensity and tone of lighting on foreground objects should balance that of the background—a brightly lit real object placed amidst dimly lit virtual objects will look unnatural. Third, some compositing processes, such as Ultimatte, can preserve shadows cast by objects in the blue room. This is extremely useful, since shadows "anchor" real objects on the virtual floor. This requires special attention to lighting, both to produce the shadows and to assure consistency with shadows cast by virtual objects (they should at least fall in more or less the same direction). Finally, set designers can place blue objects in the blue room and then "cover" them with virtual objects, allowing virtual objects to support or occlude real objects. Adding blue objects again complicates

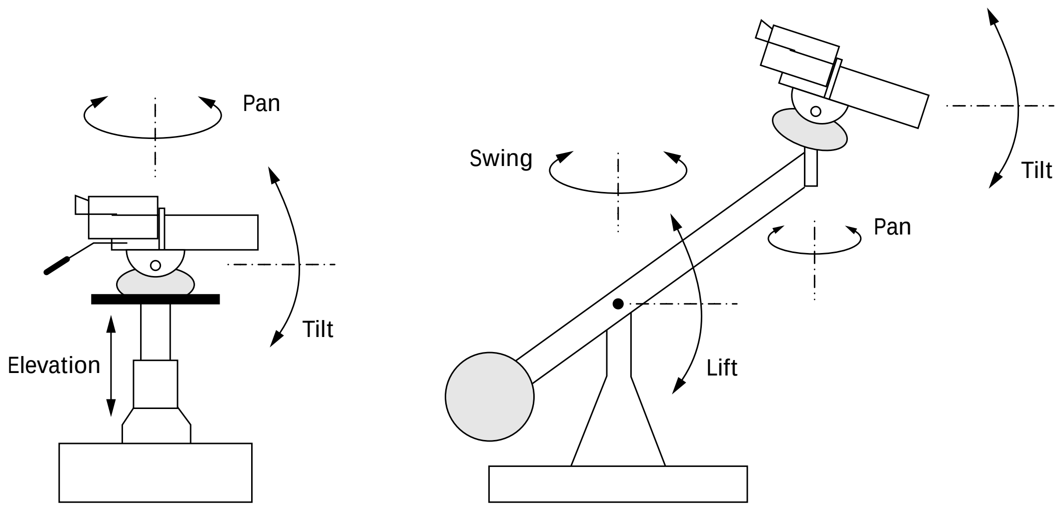


Figure 6. Camera mounts: a studio pedestal and boom.

lighting: An unwanted shadowy outline may appear around the blue object and disappear only when the compositing process is adjusted to remove shadows completely.

Camera tracking

Obviously, to render the virtual background, we need to know the real camera's position and orientation. We must also obtain its zoom setting and, if simulating depth of field on the virtual background, the focus and aperture settings as well. (In general, the aperture seldom changes and, if not tracked, can be entered manually into the virtual studio system.) The tracking systems used in virtual studios fall into two main categories: electromechanical and optical. There exists no ideal camera tracking system as of yet, and both approaches have deficiencies.

Electromechanical tracking. Developed first and still more commonly used, electromechanical systems can be active or passive. In active systems, servo-control mechanisms "drive" the camera to allow, for instance, remote control. In passive systems, sensors detect the camera operator's actions. (The sensors are often optical encoders that detect small markings on rings or other moving parts; despite this we refer to these systems as electromechanical.)

For both active and passive tracking, the system must determine the lens objective's zoom and focus ring positions plus the camera mount's various degrees of freedom. Possible mounts, with their degrees of freedom, include

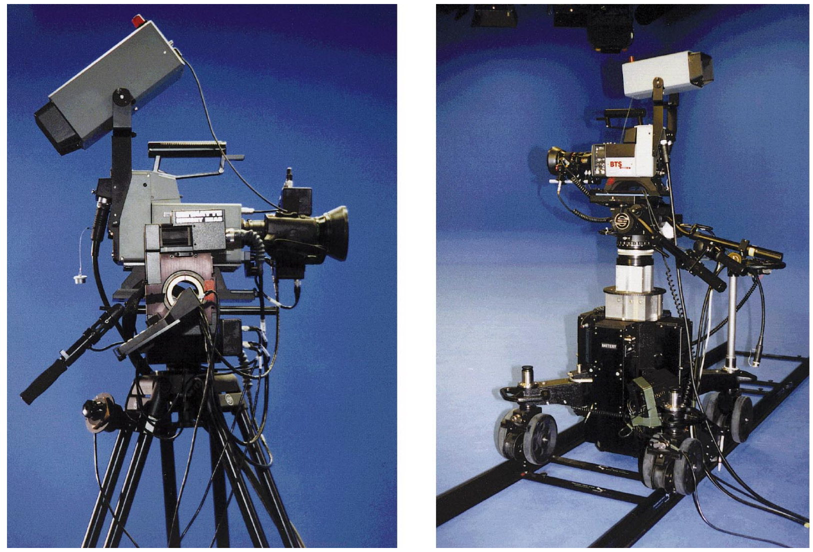
- a fixed tripod: pan and tilt
- a stationary pedestal: pan, tilt, and elevation (see Figure 6)
- a moving pedestal or dolly: pan, tilt, elevation, and displacement in one (rolling along a track) or two dimensions
- a stationary boom: pan, tilt, and boom lift and swing (see Figure 6)
- a moving boom: the above plus displacement in one or two dimensions

Electromechanical tracking can be highly accurate, providing angular measurements to thousandths of a degree and displacements to hundredths of a millimeter.[6] In addition to ruling out the use of hand-held cameras, however, it has several drawbacks:

- calibration—accurate registration requires detailed measurement of lens characteristics.
- alignment—prior to shooting, the tracking system must be aligned using potentially time-consuming procedures to determine the camera's initial position and orientation and to level the camera mount.
- backlash and vibration—the camera mount is not perfectly rigid and may bend slightly during rapid acceleration or deceleration or when under heavy load. This is particularly severe for long boom arms.

The electromagnetic tracking systems used in

Figure 7. The Ultimatte Memory Head (left) and a Thoma System with dolly and track (right).



many virtual reality applications are not suitable for virtual studio work. Aside from interference due to the proximity of lighting and video equipment, questions of accuracy and stability arise. With a video camera fully zoomed in, the horizontal field of view may be as little as five degrees. Each pixel in a video frame then covers an angle of less than 0.01 degrees, and even a slight jitter in the tracking system can shift the background by several pixels. A jumpy background is itself annoying, but when composited with a stationary foreground, the result is unacceptable. If one layer of the composite shifts unexpectedly with respect to the other, the image integrity breaks down and the viewer becomes distracted.

Optical tracking. The alternative to electromechanical tracking—optical tracking—uses pattern recognition. With visible reference points or grid lines in the blue room, image processing techniques can extract the camera's position, orientation, and field of view. Though extremely difficult to achieve in real time with the accuracy needed for virtual studio work, this eliminates the need for painstaking calibration of the lens system and can be used with any camera mount, including hand-held cameras. Optical tracking has some potential pitfalls, though:

- The tracking system may become disoriented if markers are out of focus or if an insufficient number are in view.
- The markers' real-world coordinates must be known precisely.
- Processing demands may make the tracking system lag behind camera movement by several frames, or keep operation below the video field (or even frame) rate.
- Markers in a blue room may create compositing difficulties. The markers must be distinguishable from the blue backing but should not be visible in the composite.

Tracking systems now on the market. The Ultimatte Memory Head, shown in Figure 7, was the first commercial camera tracking system sufficiently accurate for virtual studio work. This active electromechanical system measures camera pan, tilt, zoom, and focus four times per frame (supporting frame rates of 24, 25, and 30 Hz). Tracking data can be recorded onto a floppy disk residing in a control unit, sent directly to a computer over a serial line, or sent to a slave camera. In addition, the Ultimatte Memory Head can respond to tracking data coming from a second camera or a computer and can thus be used for ganging and with prerendered virtual sets (such as the configuration shown in Figure 5).

The Institut National de l'Audiovisuel (INA) in France produced an early virtual studio system called Hybrid Vision. It uses a passive electromechanical tracking system based on a Sachtler pan/tilt head with two high-resolution motion sensors and similar sensors on the zoom and focus rings. A useful audio encoder/decoder converts tracking data to an audio signal and permits recording camera motion on video tape with the accompanying camera image signal.

The Thoma System, a passive electromechanical system, consists of up to eight motion sensors. In addition to pan, tilt, zoom, and focus, the system can be configured for use with pedestal, dolly, and boom mounts (see Figure 7). It produces data at four times the video frame rate, sending each set of measurements in a 25-byte package over a high-speed serial line to the host computer.

The Radamec Virtual Reality Encoder Head is an active electromechanical system that provides pan and tilt data. An optional lens encoder provides zoom and focus. The precision is very high—the zoom and focus resolution exceeds 10,000 samples.

The Vinten TSM AutoCam, also an active electromechanical system, is a high-precision pan/tilt head with an accuracy of 36 arc-seconds (about 0.01 degrees). A lens encoder can be added for zoom and focus control.

Milo[6] is a boom and track system developed by Mark Roberts Motion Control. Its extremely rigid construction permits fast moves along several axes. This active electromechanical system tracks pan, tilt, zoom, and focus; boom lift and swing; and track displacement. (Head roll and an additional tilt-like axis are optional.)

Orad Hi-Tech Systems' optical tracking system is currently the only product of its kind on the market. It places a grid-like pattern in the blue room and passes the camera signal through a specially designed 30 gigaops-per-second digital video processor. Orad's pattern recognition software can then estimate camera position, orientation, and zoom. The video processor uses several Texas Instruments single-chip multimedia video processors.

Rendering

A virtual studio system's rendering component must produce the background layer and mask signals in sync with camera movement. If produced at the video frame rate, the background will appear jerky during fast camera moves, hence rendering should match the video field rate. For CCIR 601 video with PAL timing, the renderer must then generate a 720 × 576 image every 20 ms. Two basic approaches to background rendering exist: the first uses a 2D background image, and the second requires a 3D model.

2D background rendering. Suppose we have an image, the background source, with a known center of projection. Placing the virtual camera at this point lets us simulate pan, tilt, and zoom (and camera roll if needed) by applying a perspective transformation to a region of the background source image (see Figure 8). We can achieve the perspective change using a digital video effects (DVE) processor or a 3D graphics system:

- DVE processors perform various 2D and 3D transformations in real time; those with external control interfaces can collect the necessary transformation parameters from the camera tracking system.
- A 3D graphics system permits applying the background source image as a texture on a plane and positioning the virtual camera at the center of projection. The virtual camera is then coupled to the camera tracking system.

The background source image may be a prerendered view of a virtual set, a photograph, a video sequence (the background source need not be static, but the center of projection cannot change), or a composite of these. Thus we can have arbitrarily complex real and synthetic imagery in the background—for instance, an outdoor camera shot combined with a rendered model of a spaceship.

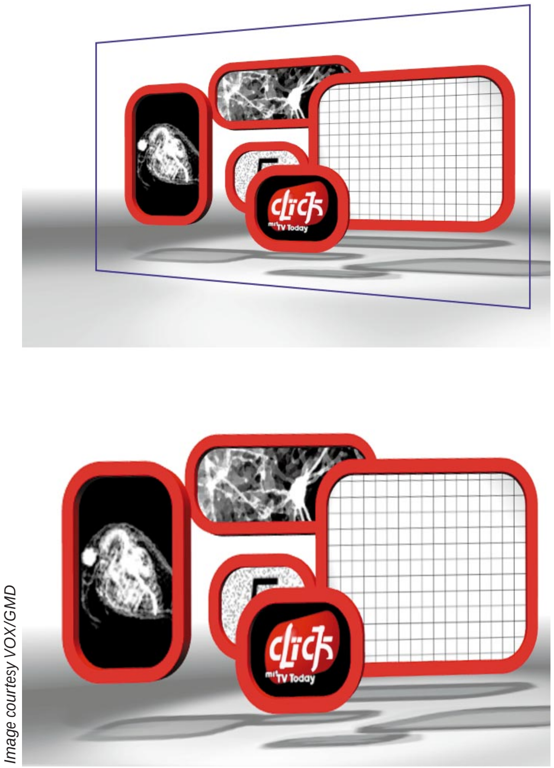


Image courtesy VOX/GMD

Figure 8. 2D background rendering: simulating a pan and zoom.

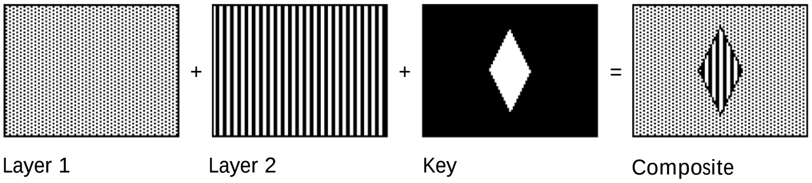


Figure 9. Compositing with a key signal.

This technique has several limitations. First, it requires some planning, since the position and orientation of the camera used to produce the background source must correspond to that of the foreground camera in the studio. Second, the field of view and resolution of the background source may limit foreground camera pan, tilt, and zoom. This poses a problem when using a DVE because the background source is limited to video resolution. Using a graphics system for background processing, we can render a very high-resolution environment map that gives full pan and tilt motion and much greater zoom range. Third, perhaps the main drawback, this technique constrains foreground camera movement—the camera must remain at the original center of projection. As we'll see later, a pan or tilt also causes a camera displacement that, unless corrected for (by similarly displacing the virtual camera), will cause a registration error between foreground and background.

3D background rendering. This approach models the virtual set in 3D and renders it with a virtual camera having the same position, orientation, and field of view as the foreground camera. There are no constraints on camera motion—the camera operator is free to move and position shots as needed (provided the motion can be tracked).

Constraints do affect model complexity. With only 20 ms available for rendering, virtual set models must be carefully designed and tuned. Models typically have a fairly low polygon count and make heavy use of textures. The virtual set designer should also consider performance optimization techniques such as spatial organization for efficient culling and multiple levels of detail. Another common technique is to pass the model through a radiosity calculation (which tends to explode the polygon count) and then apply the solution using texture maps on the original model.

Clearly, hardware support for texture mapping is essential; the hardware must also take care of antialiasing. Currently, most virtual studio systems that perform 3D rendering use either the SGI RealityEngine2 or the new InfiniteReality (although lower performance machines can be used to preview off-air cameras).

Compositing

The final stage of a virtual studio system combines the live foreground signal with the rendered background signal. The video term for this is *keying*, since an intermediate signal acts as a "key-

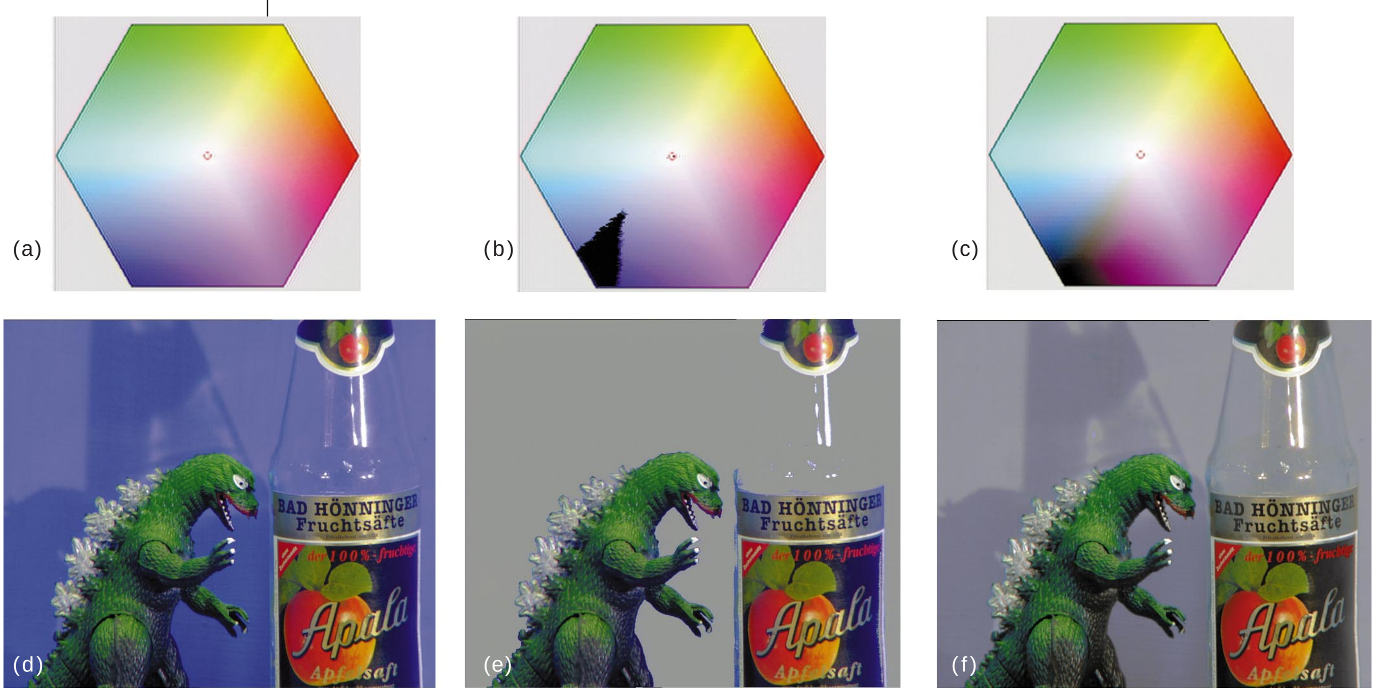


Figure 10. Compositing: (a) color wheel, (b) color wheel plus black background (chromakey), (c) color wheel plus black background (Ultimatte), (d) blue screen shot, (e) blue screen shot plus gray background (chromakey), (f) blue screen shot plus gray background (Ultimatte).

hole" and determines where one video layer shows through another (see Figure 9). Blue screen compositing derives the key signal from the foreground signal (*self-keying*). Examples of self-keying processes include *luminance keying*, where the key depends on the signal's luminance values, and *chromakeying*.

A binary key signal makes edges appear jagged and cut out, and does not support transparency effects. To remedy this, allow the key signal to vary over a range and control the mixing of the two layers, much like alpha blending. This *linear* or *additive keying* gives more natural edges and preserves fine detail but tends to leave a blue fringe around objects. It also fails to compensate for "blue spill" and gives poor results with partially transparent objects such as glass and smoke.

Ultimatte, a more complex compositing process and a near de facto standard in virtual studio systems, is not a chromakeyer, strictly speaking. Like a chromakeyer, however, it produces composites from blue screen shots. An Ultimatte keyer operates in three main steps. First, it factors out the backing color's blue level to leave the foreground black where backing was present and reduce blue elsewhere. Next, it generates a key signal based on the amount of backing color removed. Finally, it mixes the processed foreground and background together using the key signal. Ultimatte preserves foreground edge detail and can composite out-of-focus edges and semitransparent objects; it also preserves shadows cast by foreground objects onto the blue backing.

Figure 10 illustrates both chromakeying and Ultimatte. To highlight their differences, we adjusted the chromakey so that the key is nearly binary (notice how the altered region of the color wheel in Figure 10b is clearly defined and there are hard edges around foreground objects in Figure 10e).

Techniques

A successful virtual studio depends not only on powerful technology but knowing how best to use it. Here we identify some common techniques.

Camera calibration

With blue room shots, it's useful to think of "anchor points," places in the foreground that should appear to be in physical contact with the background. The feet of people standing in the blue room form anchor points, as do physical props. The absence of such points makes judging foreground object position difficult and hides slight errors in registration between the foreground and background. Anchor points amplify registration errors by drawing attention to foreground objects that appear to "slide" over the background during camera movement. The loss of continuity from such sliding is extremely distracting and generally unacceptable, making accurate registration essential to virtual studio systems. The key to accurate registration is, in turn, accurate camera calibration. Depending on the type of tracking used—electromechanical or optical—different calibration procedures apply.

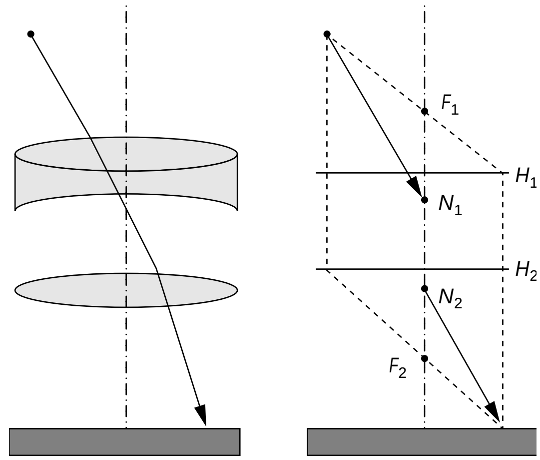


Figure 11. A lens system (left) is modeled (right) by a pair of focal points (F_1, F_2), principal planes (H_1, H_2), and nodal points (N_1, N_2). The first nodal point makes incoming rays appear as if originating from the second nodal point.

Calibration for electromechanical camera tracking. Consider an electromechanical pan/tilt head with sensors on the focus and zoom rings; assume the camera is placed on a fixed tripod. The tracking system generates a set of measurements $<m_p, m_t, m_z, m_f>$ for each video field (referring to the pan, tilt, zoom, and focus measurements, respectively). Now we must deduce the parameters of the virtual camera used to render the virtual set. At first glance, since the position of the real camera is fixed, we need only rotate the virtual camera by the pan and tilt angles and set the field of view to that indicated by the zoom measurement. However, this assumes that the real camera behaves like a pinhole camera mounted at the intersection of the pan and tilt axes, which is unlikely. In general, for cameras with moving lens assemblies, the center of projection shifts as the various lenses move.

First-order or "thin lens" optics model a lens system using six parameters: a pair of focal points (F_1, F_2), principal planes (H_1, H_2), and nodal points (N_1, N_2). The nodal points are of particular interest because if a ray is traced through the first nodal point, the lens system displaces the ray so that it emerges from the second point (see Figure 11).

Figure 12. Lens calibration targets: nodal point and field of view.

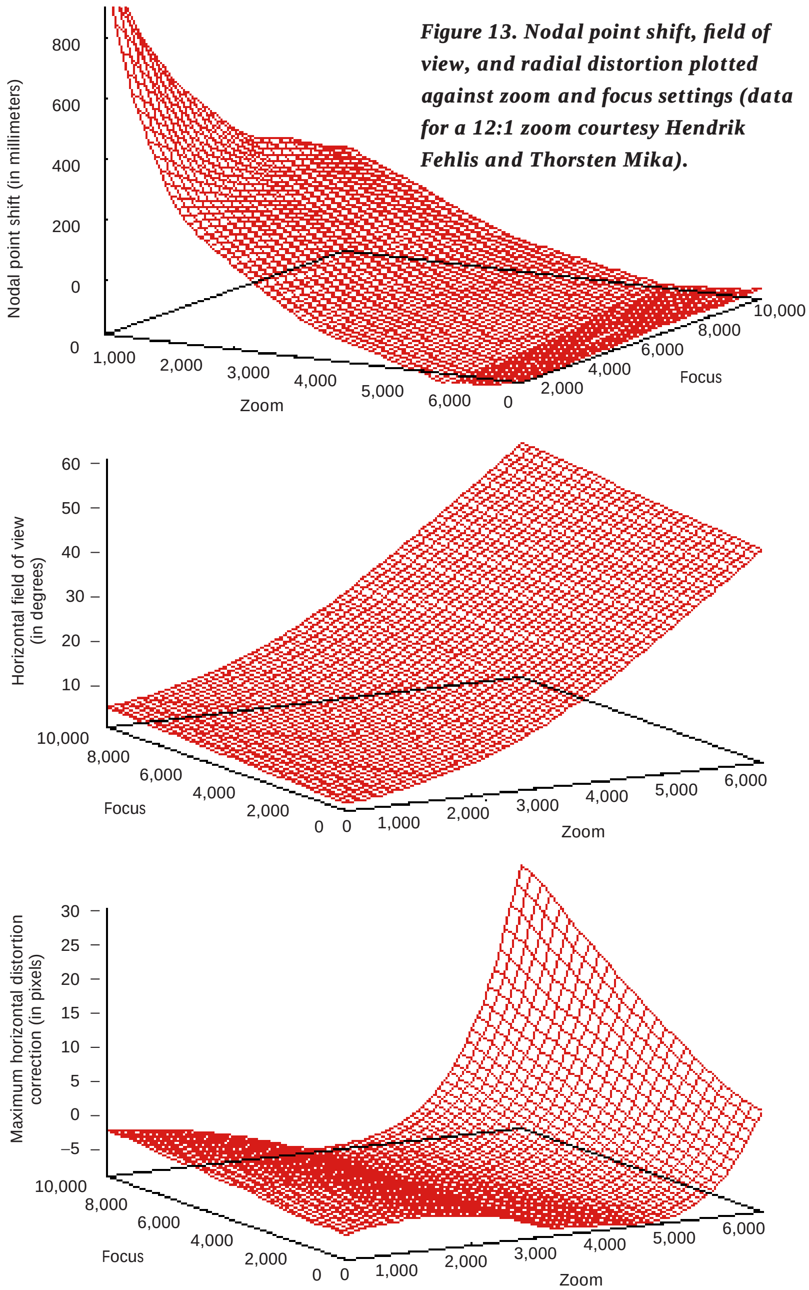


Figure 13. Nodal point shift, field of view, and radial distortion plotted against zoom and focus settings (data for a 12:1 zoom courtesy Hendrik Fehlis and Thorsten Mika).

Image courtesy of Hendrik Fehlis and Thorsten Mika

N_1 thus behaves as a center of projection on which the virtual camera must be positioned to achieve consistent registration of virtual and real images. A major difficulty is that N_1 shifts along the optical axis as the zoom and focus lenses move.

In addition to the nodal point, we need two other pieces of information for registration: the field of view (which varies with zoom and focus settings) and the first radial distortion coefficient (zoom lenses introduce considerable radial distortion, which the rendering system should correct).

To summarize, we must determine the following functions for the camera lens when using an electromechanical tracking system:

- $N_1(m_z, m_f)$: position of first nodal point (along optical axis)
- $fov(m_z, m_f)$: horizontal field of view
- $k_1(m_z, m_f)$: first-order radial distortion coefficient

These functions will vary from one lens assembly to another but can be determined offline and semi-automatically (for instance, using special targets as shown in Figure 12). Figure 13 shows some typical plots—note that the nodal point shifts by nearly one meter, and that radial distortion introduces errors of up to 30 pixels (or about four percent of frame width).

When the camera is mounted, two quantities must be measured accurately: O_h, the height of the optical axis above the tilt axis, and L_c, the distance from the pan axis to the point from which N_1 is measured (see Figure 14). Knowing these quantities and the camera mount position permits calculating the world coordinates of the nodal point for any $<m_p, m_t, m_z, m_f>$ provided by the tracking system.

Calibration for optical camera tracking. Optical tracking systems match real-world points to image points. More specifically, given a set of points for which the real-world coordinates (x, y, z) are known, the tracking system identifies the corresponding point in image coordinates (u, v). With sufficient matches it is possible to derive camera position and orientation. A popular model used for this purpose, Tsai's camera model,[7] has 11 parameters, including five internal parameters:

- f: effective focal length
- $k1$: first-order radial distortion coefficient

- *Cu, Cv*: center coordinates of the radial distortion
- *su*: scale factor

and six external parameters:

- *p, t, r*: camera pan, tilt, and roll with respect to world coordinates
- *x, y, z*: camera position in world coordinates

The model also includes several constants (such as total number of pixels in the *u* and *v* directions) that can be determined from manufacturers' specifications or measured directly.

Tsai's calibration algorithm takes a set of matched points and solves for the model parameters. Generally, an optimized solution requires at least 11 points, though if the real-world points are coplanar, then as few as five suffice. As an example, one experimental optical tracking system[8] uses eight points arranged as two coplanar groups (Figure 15).

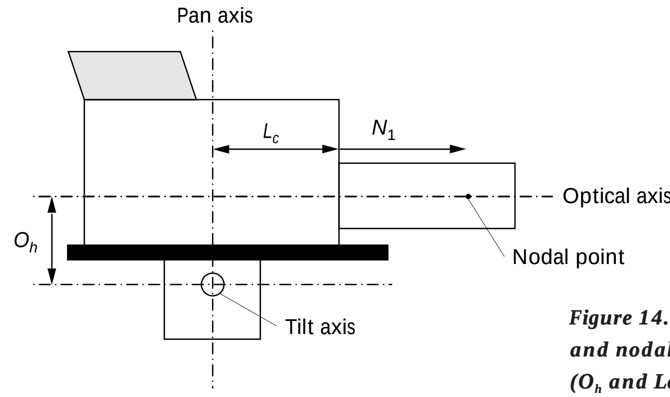


Figure 14. Optical axis and nodal point offsets (O_h and Lc).

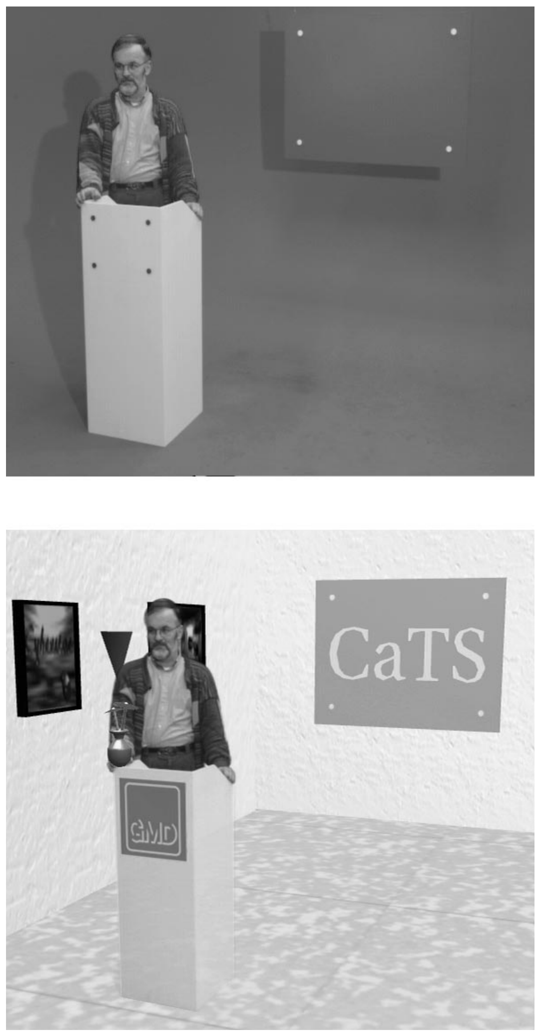


Figure 15. Optical tracking: foreground (blue room) shot on top, composite on bottom. Eight reference points are used, four on the podium and four on the hanging blue frame.

Synchronization

Figure 16 (next page) shows a timing diagram for a virtual studio system. Note that all components are "genlocked" to a master sync signal (provided by a video sync pulse generator). Genlocking is useful for several reasons. First, video mixing equipment such as a chromakeyer or Ultimatte requires that the foreground and background signals be in video sync. Second, the master sync provides a common time base for all components to simplify detecting and correcting timing problems.

In Figure 16, video fields are numbered sequentially, so fg_0 indicates the first foreground field. Assume that the tracking system produces data at the field rate; these measurements are indicated by the sequence of m_i. Also assume that the renderer is double-buffered and renders at the field rate. It will take some time for a measurement to be transmitted to the renderer and the background to be rendered, hence bg_0 is not placed in the front frame buffer until field period 1.

Next the renderer's output (located in the front frame buffer) is scan converted to a CCIR 601 signal; assume this takes two field periods. The final result is that bg_0 is not composited with fg_0 but with fg_3, or, in other words, the background is delayed by three fields—for PAL, a delay of 60 ms.

Visually, delay in a virtual studio system manifests itself as loss of registration during camera moves. For instance, if the operator pans the camera and comes to a rapid halt, the viewer sees a stationary foreground while the background continues to pan for a few fields. The delay can be hidden by low camera acceleration or decelera-

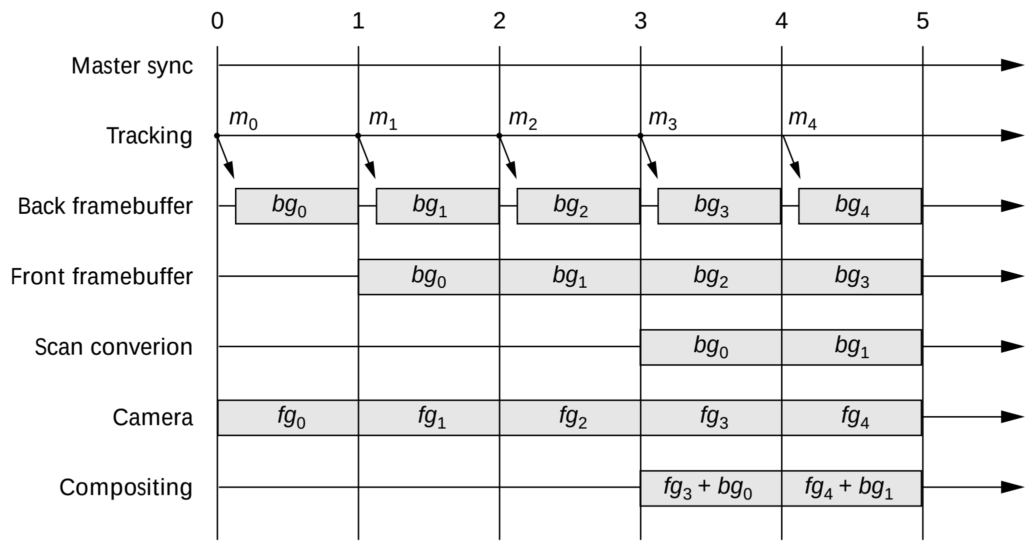


Figure 16. A virtual studio timing diagram. Time proceeds left to right, and the horizontal intervals denote field periods.

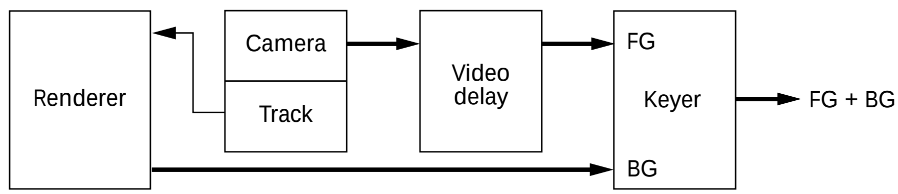


Figure 17. A virtual studio system configured with a foreground delay.

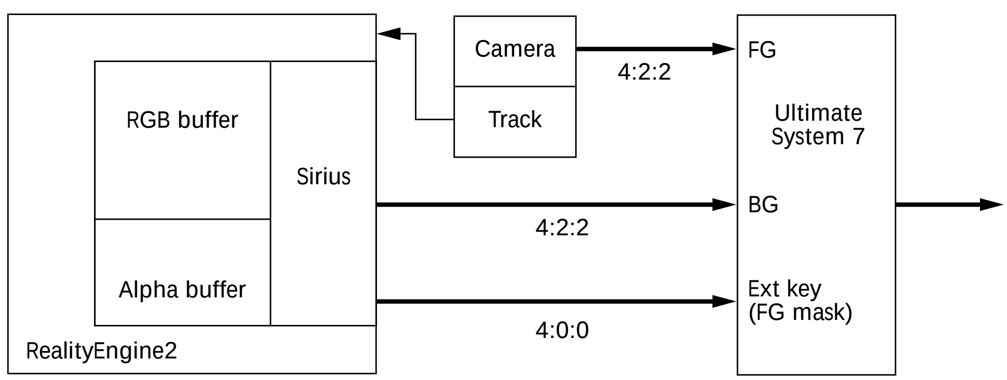


Figure 18. A virtual studio system configured with a foreground mask.

tion, but this constrains the camera operator. In general, virtual studio systems must compensate for background delay. Some possible techniques include

- Advancing the tracking data. Apply predictive filtering to tracking measurements and extrapolate the data forward in time.

- Delaying the camera head. Some active electro-mechanical tracking systems can be programmed to delay before acting on move requests (though the move would be passed on to the rendering system without delay).

- Delaying the foreground. A video delay line can be introduced between the camera and the compositing system (see Figure 17).

A correctly compensated delay, no matter how long, is not noticeable to someone viewing only the final composite. It may, however, still be noticeable to people in the studio. For example, the video delay line, the simplest to implement, requires a matching audio delay to maintain lip sync. A delay of more than a few frames can be extremely distracting to an actor should he hear both his own voice and the delayed signal. Similarly, the camera operator, observing the delayed composite signal in the viewfinder, will notice that the image does not respond immediately to camera movements.

Advancing tracking data or delaying the camera head also present problems, in particular when the delay is very long (four frames or more). Advancing tracking data is error-prone and may be visible during abrupt camera movement, while adding a delay to the camera head can make the camera difficult to handle.

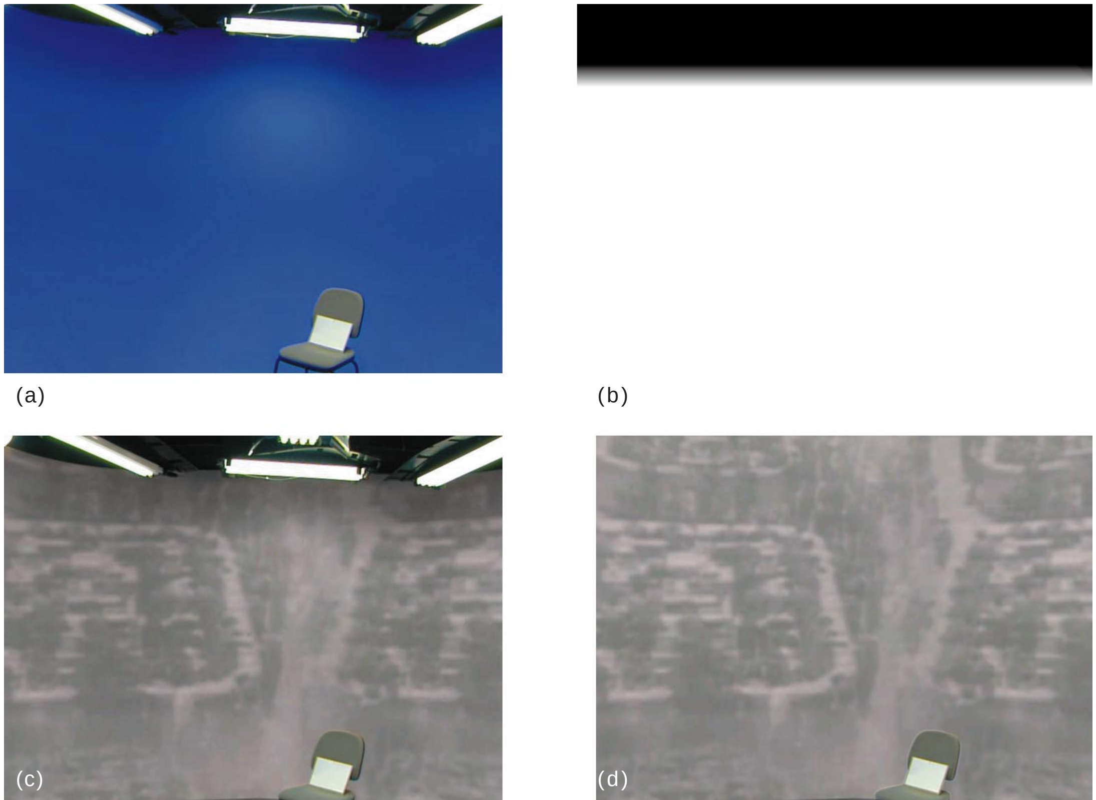


Figure 19. Garbage masks: a) foreground, b) garbage mask, c) foreground+ background, d) foreground+ background using garbage mask.

Real-time effects

A main strength of virtual studio systems is that they are "live": rendering and compositing are done in real time rather than offline or during postproduction. Virtual studios can perform many animations and special effects seen with offline systems—for example, props can move and change their size or morph from one form to another. Such animation is manually triggered and controlled or, in a tightly scripted production, triggered by some form of time code. Since these effects must be performed in real time, the modeler must be aware of their performance impact; thus virtual studio effects are very "device dependent." Here we discuss special effects that rely on features of the rendering and compositing hardware, in particular the commonly used Ultimatte System 7 (or System 8), the SGI RealityEngine2 (or InfiniteReality), and the SGI Sirius video board.

Foreground masks. The Ultimatte, like many chromakeyers, accepts an external key that is combined with the key generated from the foreground signal. The external key identifies background regions that should appear over the foreground, hence it masks out the foreground. As an example, Figure 2 shows a globe (a background element) forced in front of the blue room (foreground) signal. This is easy to implement and forms the basis for effects such as people moving behind virtual walls or furniture. Using a RealityEngine2, the foreground mask is typically rendered in the alpha buffer, then output by the Sirius board in 4:0:0 format and sent directly to the keyer (see Figure 18).

Garbage masks. Suppose the blue room has a low ceiling or is too narrow for wide-angle shots. Knowing the geometry of the blue room, the position, orientation, and field of view of the camera permits generating a "garbage mask" that forces the real ceiling (or side walls) of the blue room to be keyed out. This can be accomplished by modeling a ceiling for the virtual set and creating a foreground mask, generating a horizontal band in the alpha buffer, or controlling a keying window in the compositing device (the latter two can be used even if the virtual set has no ceiling). Figure 19 shows an example of the second method.

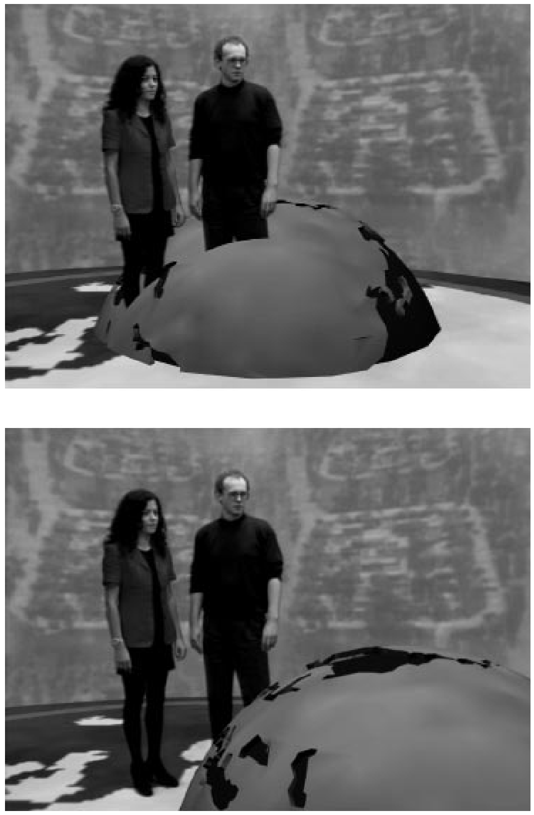


Figure 20. Z-mixing with poorly chosen foreground z (top), and with foreground mask (bottom).

Image courtesy WDR/GMD

Figure 21. Video texture with transparency.

Figure 22. Sitting on virtual steps.

Z-mixing. A *z*-mixer combines two video sequences based on their depth values. Obtaining the depth information proves difficult. VMix,[9] a hardware device for *z*-mixing, takes the renderer's *z*-buffer as a per-pixel source of depth values for the background; the foreground is assigned a constant *z* value. For a constant foreground *z* value, we can also simulate *z*-mixing by rendering a foreground mask with the far clipping plane set to the foreground *z* value, which is entered either manually or via a tracking system. For instance, Orad fixes a small video camera above the blue room and determines a representative foreground *z* value by identifying objects in the blue room. How the foreground *z* value is obtained is important, since many values produce bizarre results (note how the people disappear into the globe in the left image of Figure 20).

Video textures. The Sirius board supports full-frame-rate video textures, a powerful feature. The texture source can be any video device—usually a video tape or hard disk recorder, although for teleconferencing in the virtual studio, the source could be a live camera satellite feed. The video texture may appear as a simple flat "video wall" in the virtual set or, like other textures, be mapped onto more complex geometry. If the video texture has an accompanying key signal, the key can be treated as an alpha channel and parts of the video texture can be made transparent. For instance, the announcer appearing in the pyramid in Figure 21 is actually a video texture with an alpha channel created by a chromakeyer.

Blue props. Placing blue objects of various shapes and sizes in the blue room lets real objects "interact" with virtual objects. For instance, a blue box aligned with a virtual table permits placing real objects on or behind the table; or several blue boxes on the floor might allow an actor to sit on the steps of a virtual staircase (see Figure 22). Using blue blinds lets actors disappear behind virtual walls and walk through virtual doorways. Although masks can achieve the same effect, blinds have the advantage that real shadows on their surfaces are preserved through compositing and will appear correctly placed on the corresponding virtual surfaces.

Virtual actors. Real-time animated characters—virtual actors—provide another form of interaction between real and virtual objects. A virtual actor can respond to real actors' movements and gestures in the blue room, and can carry on conversations with real actors or other virtual actors. One or more operators typically control a virtual actor's movements using an input device such as a body or facial motion capture system, a dataglove, or simply a joystick. One operator also provides the virtual actor's voice. In a live broadcast, remote viewers can be put in control of a virtual actor, for instance using telephone touch tones on a phone-in show to "steer" the virtual actor about the set.

Table 2. Commercial virtual studio systems.

System	Company	Tracking	Rendering
3DK http://viswiz.gmd.de/DML/vst/vst.html	GMD	Electromechanical	3D
Best http://www.studio.sgi.com/Features/VirtualSets/brainstorm.html	Brainstorm	Electromechanical	3D
Cyberset I/O Cyberset D/E http://www.orad.co.il	Orad	Optical	3D 2D
Elset http://www.studio.sgi.com/Features/VirtualSets/accom.html	Accom	Electromechanical	3D
Hybrid Vision	INA	Electromechanical	3D
Larus, Otus http://www.rtset.co.il	RT-Set, Chyron	Electromechanical	3D
MindSet http://www.es.com/PR/pr04-15-96.html	Evans and Sutherland	Optical	3D
Reality Tracking	ElectroGIG	Electromechanical	3D
Synthevision	NHK	Electromechanical	2D (Hi-Vision background)
Vapour http://www.discreet.com	Discreet Logic	Electromechanical	3D
VideoScape	Vinten	Optical, electromechanical	Ganged cameras
Virtual Scenario	Radamec	Electromechanical	2D
Virtual Theater http://www.softimage.com	Softimage	Electromechanical	3D

Shadows. A final challenge is adding shadows to virtual sets. Here we differentiate between real shadows on real objects, real shadows on virtual objects, virtual shadows on virtual objects, and virtual shadows on real objects. The first happens naturally and depends on studio lighting; the third is a traditional rendering problem. Unique to virtual sets are techniques for casting real shadows on virtual objects and virtual shadows on real objects. As mentioned earlier, compositing hardware can generate a key that is only partially transparent in regions where shadows darken the blue room walls or floors. This also darkens the corresponding areas of the background to create the appearance of real shadows falling on the virtual background. However, this method requires that real and virtual geometry coincide and would not give correct results if, for example, the blue room floor is empty while the virtual floor is populated with objects.

For real shadows to actually wrap around virtual objects, requires multipass rendering techniques such as "projecting" the key signal as an additional texture onto the virtual set. Techniques for casting virtual shadows on real objects are still experimental: possible approaches include physically projecting a rendered signal onto the blue room and its occupants, or calculating a depth map for the camera signal and providing this to a *z*-mixer.

Commercial systems

Table 2 lists the virtual studio systems now on the market (most Web page references also appear at the Virtual Studio Home Page, http://viswiz.gmd.de/DML/vstudioHome.html). In addition to the system and company names, the table indicates whether electromechanical or optical tracking is used, and whether background rendering is 2D or 3D. Although most systems currently use electromechanical tracking, integrating optical tracking proves fairly simple. Also note that the 3D systems can perform 2D rendering if needed.

The commercial 3D virtual studio systems show fairly similar functionality and performance (assuming similar hardware) but differ in accuracy of tracking and registration, ease of use, and general robustness. Rather than attempt to evaluate them (difficult to do without actually running them side by side), we'll simply list some of their more useful features:

- support for multiple cameras
- preview of "off-air" cameras using less expensive rendering hardware
- interfaces for several tracking systems
- acquisition of lens calibration data
- simple camera alignment procedures
- multiple model import formats
- high-resolution textures
- real-time antialiasing and texture mapping
- real-time effects
- export of tracking data for offline rendering

The preview capability refers to a multicamera configuration where low-end hardware renders the background for off-air cameras and a high-end system runs with the on-air camera. This decreases the cost of a multicamera system and is fairly easy to implement (though switching must be properly timed to lessen the danger of the new camera being keyed with the old background for a frame or two after a cut). When setting up an electromechanical tracking system, alignment involves locating the camera in the virtual set's world coordinates (or the blue room's coordinate system). Any automatic or semi-automatic procedure (such as placing a target of known dimensions at a reference point in the blue room) will save much time during shooting.

Discussion

Perhaps because of virtual studios' novelty, many people either overlook or do not clearly state their limitations. We therefore discuss some concerns we have for the current generation, at least, of virtual studio systems.

First, production personnel must be aware of the constraints virtual studio technology imposes. The most obvious is the need to shoot in a blue room. Blue rooms generally have bad acoustics (from the bare walls and curved corners), and actors find it difficult to position themselves or interact with invisible objects. Also, blue room lighting is difficult, and once the lighting is set, changing it (perhaps for some dramatic effect) poses yet another challenge.

Another constraint stems from virtual set complexity. The rendering system limits such things as polygon counts, texture sizes, and lighting effects. The set designer, and any other people involved with constructing the virtual set, must know the limitations of the rendering system.

Second, virtual studio systems demand new skills for successful operation. Camera operators must understand the tracking system's capabilities and be comfortable with alignment procedures. Actors need training in blue room techniques, and directors must be aware of the overall complexity of operation and the functioning of the various subsystems for rendering, tracking, and compositing.

Third, start-up costs, rapid obsolescence, and integration with existing studio equipment make it difficult to decide when best to acquire virtual studio technology. This is especially true in light of the broadcast industry's current restructuring.

Virtual studio systems have many advantages, however. Many users note economic benefits such as reduced storage costs (virtual sets do not take up much physical space), reduced construction costs (fewer expensive props or materials, easy to set up and dismantle), reduced studio costs (one blue room can be "configured" for different productions simply by switching virtual sets), and reduced design costs (easy to change a wall's color or floor's texture).

Virtual studios also open up new creative possibilities in set design and visual effects; they may reduce production time and allow greater experimentation. For instance, since virtual studios operate in real time, those involved in a production can see results immediately without waiting through possibly extensive postproduction delays. Virtual studios may also lead to new forms of audience interaction and participation: Because we can now place actors into a virtual environment in real time, viewers also may soon be able to enter the virtual world.

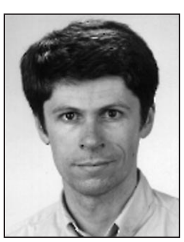
Virtual studio systems will continue improving technically as rendering and tracking systems become more advanced. More flexible compositing is also likely. Exactly how real studios will change to accommodate the virtual studio remains unknown, but this technology promises to play a significant role in the future of broadcasting production. MM

Acknowledgments

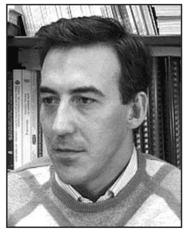
We would like to thank our colleagues in the Visualization and Media Systems Design group and members of Horz & Schnepf Co. for their assistance in developing the virtual studio project at GMD.

References

1. S. Shimoda, M. Hayashi, and Y. Kanatsugu, "New Chromakey Imaging Technique with Hi-Vision Background," *IEEE Trans. on Broadcasting*, Vol. 35, No. 4, Dec. 1989, pp. 357-361.
2. "BBC News Dives into Vertigo Reality," *Int'l Broadcasting*, April 1993, p. 9.
3. L. Blondé, "The Mona Lisa Project," *Proc. European Workshop on Combining Real and Synthetic Image Processing for Broadcast and Video Production*, Springer Verlag, Berlin, 1994, pp. 1-9.
4. L. Blondé et al.,"A Virtual Studio for Live Broadcasting: The Mona Lisa Project," *IEEE MultiMedia*, Vol. 3, No. 2 , Summer 1996, pp. 18-29.
5. D. Hughes, "Virtual Studio—Ultimatte 8," *IRT Symp. on Virtual Studio Technique*, IRT, Munich, 1996.
6. *Milo Technical Specifications*, Mark Roberts Motion Control Ltd., 1996.
7. R. Tsai, "A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses," *IEEE J. Robotics and Automation*, Vol. RA-3, No. 4, Aug. 1987, pp. 323-344.
8. P. Wiβkirchen, K. Kansy, and G. Schmittgen, "Integrating Graphics Into Video Image-Based Camera Tracking and Filtering," *Proc. 3rd Eurographics Workshop on Virtual Environments*, Springer Computer Science, Berlin, 1996, pp. 74-84.
9. W. Schmidt, "Real-Time Mixing of Live Action and Synthetic Backgrounds Based on Depth Values," *Proc. European Workshop on Combining Real and Synthetic Image Processing for Broadcast and Video Production*, Springer Verlag, Berlin, 1994, pp. 23-31.



Simon Gibbs is a staff engineer with Sony Distributed Systems Laboratory in Santa Clara, California. Before joining Sony he was a senior scientist at GMD (the German National Research Center for Information Technology), where he was responsible for the Digital Media Lab within the Institute for Media Communication. Gibbs earned a PhD in computer science from the University of Toronto in 1983. Research interests include office information systems, computer-supported collaborative work (CSCW), object-oriented programming, and multimedia systems.



Constantin Arapis obtained an MS and a PhD in computer science from the University of Geneva in 1986 and 1992, respectively. From 1994 to 1996 he was research scientist at GMD (the German National Research Center for Information Technology). Since September 1996 he has been Maitre d'enseignement et de Recherche at the University of Geneva. Current research interests include tele-education applications, teleconferencing systems, and Internet multimedia applications.

Christian Breiteneder is an associate professor at the Institute of Applied Computer Science and Information Systems at the University of Vienna, Austria. He received a Dipl. Eng. in computer science from Johannes Kepler University, Linz, Austria, 1978 and a PhD in computer science from the University of Technology, Vienna, in 1991. His current interests include multimedia information systems, media processing systems, and telepresence systems. Breiteneder is a member of ACM, IEEE, and IEEE Computer Society.

Vali Lalioti is a research scientist with the Visualization and Media Systems Design group at GMD, Germany. She received her BS and MS from the University of Crete, Greece, and a PhD in computation from the University of Manchester Institute of Science and Technology, UK. Her main research interests include virtual studios and collaborative virtual reality environments.

Contact Gibbs at Sony Distributed Systems Laboratory, 2350 Mission College Blvd., Santa Clara, CA 95054, email simon@arch.sel.sony.com.