Editors: Theresa-Marie Rhyne and Lloyd Treinish

Visualizing the Real World

Lars Nyland and Anselmo Lastra

University of North Carolina at Chapel Hill

What is visualization of the real world? In our view, it means visualizing the environments around us, at scales we're accustomed to, sensing what we would normally sense. Unlike others in the field of visualization, we don't aim to provide x-ray vision, heat or airflow models, or small-scale measurements (such as the inspection of parts to the nearest micron or determination of molecular structures by x-ray crystallography). And neither do we aim for larger-than-life views, such as those presented by geographic information systems (GIS) of our world or the models of worlds beyond provided by astrophysicists. Instead, we focus on what humans currently sense, at scales that we navigate daily, to record and share experiences.

Imagine being able to visit the most wondrous places on Earth, whether it's the top of Mt. Everest or the Grand Canyon. Or the destination could be a place of substantial cultural, historical, or religious significance, such as the Parthenon, the Great Wall of China, or the Taj Mahal. The scene could represent artistic culture, such as any of Gaudi's architectural creations in Barcelona, the annual Ice Hotel in Sweden, or the Guggenheim Museum in New York. What will it take to give the illusion that you are there?

Capturing cultural heritage sites worldwide allows not only the sharing of cultural artifacts but also provides a mechanism for preservation in the event of destruction, whether planned or accidental. The Digital Michelangelo,[1] the Great Buddha of Kamakura[2] and the Florentine Pièta[3] projects are leading examples of such preservation, digitizing objects and environments of significant cultural heritage. We're involved in a pilot project with the Thomas Jefferson Foundation to create detailed models of Monticello. Figure 1 shows a rendering of Jefferson's library. These are only a few of the possibilities for this technology. We cite other uses for this technology in the "Using 3D Real-World Scenes" sidebar.

Acquisition methods

The ideal scene-capture instrument would densely sample all the environment's surfaces, automatically and instantaneously. A close approximation for static scenes (with static illumination, thus indoors) would be a robotic platform that purposefully traverses the environment and returns with the complete geometry and view-dependent color. We can dream of an instrument that also captures data along the time dimension so that, for example, we can understand the dynamics of automobile crashes or enjoy basketball games.

Of course, this luxury doesn't exist yet. It takes time and careful planning to acquire geometry, surface coloring, lighting, and the view-dependent effects of all surfaces in the scene. These goals are difficult to achieve completely, but are nevertheless pursued with a wide assortment of techniques and technologies.

Many methods exist for capturing scene geometry. The simplest, but perhaps most widely used, is the tape measure, which sparsely captures the relative geometry of a scene, such as a car accident (skid mark lengths, positions of cars, positions of occluders). Surveyors use theodolites and global positioning to determine the exact locations of significant artifacts in the scene. For higher density, we can use methods from computer vision (stereo, structured light, and so on) and photogrammetry to determine geometry while also capturing the color

1 Rendering of Thomas Jefferson's library at Monticello using fixed-size points. We acquired the 3D data from a single point of view with a DeltaSphere-3000 laser range finder and a high-resolution digital camera.

Using 3D Real-World Scenes

Besides documenting cultural heritage, the following are examples of how capturing 3D real-world scenes might be of great interest:

- *Forensic science*. We can capture crime and accident scenes and digitally preserve them, replacing or supplanting the photographic records and the extensive environments currently stored for later legal use. Realistic color is quite important here, attested to by the reliance on photographs in current forensic procedures. Figure A shows a staged car crash we captured.
- *Entertainment industry*. Movies, television, and gaming could readily adopt digitized real-world environments. In fact, environment and object capture is already in limited use for special effects in movies. We expect more use in the movie industry soon, as well as game environments and digitized game characters and players.
- *Surveying and engineering*. This is a less glamorous area, but one with huge demand, where quick acquisition of comprehensive, as-built models has a substantial impact. Retrofitting processing plants, upgrading offshore oil drilling platforms, and understanding the available spaces in submarines are just a few processes that can benefit from digitized environments.
- *Teleconferencing*. Today's technology doesn't typically lead to a satisfying interaction among teleconferencing participants (just note how many people travel to spend time with one another, even though it's expensive and time consuming). There's an indication that the big difference between a face-to-face visit and a teleconference (using television) is the lack of 3D. Participants don't become involved, and it's difficult to judge participants' reactions and emotions with current telecollaboration technologies. Some human–computer interaction (HCI) experts believe that the ability to share 3D environments and 3D models in real time will drastically improve telecollaboration. Figure B shows work that aims to provide 3D real-time telecollaboration technology.

A View from a 3D model of an automobile accident (staged) to illustrate forensic science visualization. (Automobiles courtesy of Krause and England.)

B A telecollaboration scenario lets a person look through a "hole" in the wall into one or more remote spaces. The viewer's head is tracked and the 3D data is projected in stereo. Multiple cameras create a 3D model of the remote participants several times per second, and they compose it with the previously acquired static background. (Image courtesy of Henry Fuchs, Herman Towles, and the National Teleimmersion Initiative. Photo by Larry Ketchum.)

imagery. More precise methods of environmental geometry determination rely on scanning laser rangefinders, such as the one we have assembled. There have even been demonstrations of range imagers that capture a full frame in a fraction of a second, yielding multiple range images per second (as high as 100,000 frames per second, given the right conditions and sensors).

Acquiring high-quality color (and rendering it later) to place on the geometry is a difficult problem, since our eyes have such high acuity and a wide range of sensitivity. Cameras, whether electronic or film-based, can't come close to matching the wide dynamic range we perceive. One solution is to acquire high-dynamic range images, but even when we acquire them, we have the related problem of rendering them.

Modeling the view dependence of surfaces in a typical scene requires color samples from many different locations (the number depending on the materials), with very dense sampling required for glossy surfaces. If we wish to enable movement of the lights when rendering the resulting model, we need to measure the surfaces' material properties, a difficult task even under laboratory conditions.

Assuming some mechanism for capturing geometry, color, and surface effects, the next problem becomes accurately mapping the data sets to each other. We rely on accurate construction coupled with calibration techniques to understand and eliminate the distortions from our system.

In any interesting 3D environment, it's usually impossible to see the entire environment from a single viewpoint. Historic churches and chapels are filled with

pews, columns, podiums, and doors. Accident scenes contain objects that have fallen or collided. Sculptures contain self-occluding structures, like arms and legs or multiple figures that hide other parts of the sculpture. We usually fill in these hidden areas by obtaining data from as many viewpoints as needed to adequately represent the surfaces. We believe that capturing every surface is a difficult, if not impossible problem, because of the large number of surfaces obscured from any viewpoint or surfaces that are difficult to measure. Fortunately, many surfaces aren't interesting, such as the undersides of chairs or desks. Also, some places are too small to accommodate the acquisition equipment.

To solve the problem of occluded areas, we acquire data from many locations, that are then registered and merged into a single model. We can instrument the acquisition process, creating a model that's correct by construction, or register afterwards, since a large percentage of any part of a scene is viewable from many locations. Furthermore, planning where to position acquisition devices to take as few views as possible (the next-best-view problem) is difficult even when we know all the geometry. In an acquisition scenario, nothing is known until data acquisition begins, thus we must solve the problem incrementally.

Acquisition difficulties

All measurement technologies have their weaknesses. Light-based methods (lasers, structured light, and so on) all fail on reflective, dark, and transparent materials. For accurate models of people, hair is a notorious problem because it's shiny, refractive, and often dark. Vision-based methods have trouble on smooth areas, such as ordinary walls, carpeting, or the sky. Contact methods—such as computer-controlled measuring machines or the simple tape measure—often take too long or sample the scene too sparsely.

In short, it's apparent that human mental models of scenes are heavily filled in from our earlier life experiences. It takes no time at all for anyone to understand the shape and beauty of ice sculptures, yet it's difficult to imagine any method of acquiring an accurate model of such an object without substantial modifications to the scene.

Model building

Acquiring an environment produces millions or billions of colored points in space, perhaps with view-dependent information attached. Some in the image-based rendering community might argue that this is all we need, at least for now. However, rendering directly from the huge amount of raw data is currently impossible. Instead, we must put the data into a form that suitably represents the scene where it can be processed and rendered. This process typically includes registration of the multiple views, merging to cull redundant data, followed by tessellation and simplification.

For example, in scanning just the library and adjoining annex at Monticello (about 200 square feet), we took nine panoramas of range and color, accumulating more than 100 million range samples and 1,000 color images. For this small project of just two rooms, we used approximately 4 Gbytes of disk space. While disk space may be inexpensive, it's nearly impossible to process or simply render these data, let alone data for the entire house. The entire estate would undoubtedly consume more than 100 times as much storage, requiring drastic measures to allow interactive rendering.

We can currently acquire more data than we can possibly use (for any of the mentioned technologies and domains), and yet we never have quite enough data to build an entire scene. The problems come from inadequate view planning and inadequate measuring hardware.

Most range acquisition systems make measurements independent of the scene complexity. This is in direct contrast to our eyes, where most of the acuity is centered at the fovea, and only specialized sensitivity exists away from the fovea (such as motion detection). Range scanners acquire data just as densely on large, flat walls as they do on interesting sculptures, bookshelves, automobile grills or other intricate surfaces. Decimation, therefore, occurs afterward, and is usually some well-explicated form of compression, such as decimation by surface curvature, that drastically reduces the data complexity (without altering the model too much). This is an area where texture mapping provides an excellent reduction in complexity, since we can simplify the geometry substantially while keeping the textures as detailed as possible. Surprisingly, few people use this, often eliminating the color with the geometry, since the large number of geometry samples often overwhelms the computational resources.

Simplification often achieves substantial reduction in complexity, but there's room for improvement. Walls are perhaps the best example, since they're easily modeled with polygons. The tessellation systems do a good job, taking millions of samples (with potentially two triangles for each sample) and reducing them to thousands of triangles. But rarely is a rectangular, planar area simplified to just two triangles. Although we can reduce the data by several orders of magnitude, it appears that there may be a few more to go.

Despite these problems, we proceed onward, taking on the tasks of registering of data taken from multiple points of view into a single, unified coordinate system; registering of the multiple data sets (geometry, color, surface properties); and simplifying each sample's information content (geometry and color).

Rendering

Our primary goal is interactive visualization of the acquired 3D data models. We have explored conventional rendering techniques and the newer sample-based approaches to attack this problem.

The data sets' sizes make model simplification and model management imperative. When we use head-mounted displays, we require fast rendering, and the only way to achieve this is to drastically simplify the range and color information (even when using the newest rendering hardware).[4] While this approach works well on static data sets, tessellation and simplification are problematic when rendering live data.

When rendering patches smaller than a pixel, does it make sense to use triangles as the basic rendering primitive? That's the question that drives us to reexamine the

types of primitives that we use for rendering. Figure 2 shows an example of a plant rendered using acquired samples. The Plenoptic Modeling method[5] reprojects samples from images that have depth, along with color, stored at each pixel. The biggest problems with these depth images are that holes or gaps typically appear when we reproject the image and the sampling rate can vary considerably across the final image. As an example of the latter, imagine that we sample a plane surface at an oblique angle but view it from a direction normal to the surface. The reconstruction will likely suffer from a lack of samples. The opposite can also occur when we oversample the surface relative to the desired viewpoint, a situation that requires careful filtering to yield a high-quality result.

A solution to both the missing-surface and the sampling-rate problems is to combine samples from multiple views to render the final image. The Layered Depth Image (LDI) Tree[6] is a hierarchical data structure that enables fast access to samples at the resolution appropriate for the rendering viewpoint. Another sample-based approach, QSplat[7] from Stanford University, is designed for fast, progressive rendering of large, sampled models.

We've also explored special-purpose hardware for rendering from depth-image samples. The intent is to take advantage of the nature of the sampled models—a large number of dense sample points, each of which is projected to a small area on the screen, typically one pixel in size. The WarpEngine architecture[8] renders from 16 × 16 sample depth images. Unlike conventional scan conversion, samples are rasterized in a forward direction, without regard for whether they land on the center of a pixel. To correct for the reconstruction errors that would result, we save subsample offsets, one or two extra bits for each of x and y, and use this more precise location for high-quality filtering to obtain the final image. A high degree of parallelism in the WarpEngine architecture results from the coherence inherent in the depth images used as primitives. Figure 3 shows rendering from our WarpEngine simulator (using real-world data).

Most of the sample-based approaches haven't rendered view-dependent color (the notable exceptions being the Light Field and Lumigraph). Recent work[9,10] uses compressed representations to store the color from a number of images at surface points. Using multipass methods, they can achieve interactive rendering rates. The color-to-geometry mapping problem remains difficult, however.

The big problems

Time is the biggest enemy in capturing scenes of the real world. Most acquisition techniques involve scanning hardware and/or hundreds of photographs, requiring from a few minutes to several hours to capture data from a single point of view. Multiple points of view magnify the time-consuming nature of the process. Even if the data from a single viewpoint could be captured instantly, acquisition from multiple locations would still take time (or the use of multiple acquisition devices, but they may then appear in each other's views). Going outdoors is problematic because there's almost always a breeze and the sun is always moving.

2 Plant rendered using splatting instead of conventional polygonal rendering. The 3D data, taken with a laser range finder, came from multiple acquisition views to minimize holes.

3 Rendering from our WarpEngine hardware simulator showing high-quality rendering of the reading room in our department. We expect that the WarpEngine, a system for directly rendering from sample-based data, will render HDTV resolution images at 60 Hz.

Thus far, most real-world scene digitization has taken place where scenes are static and lighting is controlled. One highly visible alternative to the problem of long acquisition is the Timetrack work of Dayton Taylor, seen on commercial television and movies as stopped or slowly moving scenes with rapid camera movement. He solves the time and multiple view problems by using 40 to 80 carefully placed cameras shuttered simultaneously (or with very slight delay) around a precisely choreographed event. Even though the result is only a sequence of images, it's extremely effective at conveying 3D to the viewer. Visit http://www.timetrackvfx.com for explanations and samples of Timetrack technology from around the world.

We're uncertain how to model all the data in a scene. Earlier we discussed polygonal and image-based representations and mentioned some of the problems of each. All known representations suffer from one or more of the following shortcomings: slow rendering, large demands on storage, inability to represent view-dependent color, or tedious preprocessing.

Sharing data is also a problem. To convey the significance of our work (the community, not just us), we try to find compelling scenes for acquisition. Unfortunately, along with these data come restrictions that restrict data

4 A detailed wireframe rendering of a pressure gauge and pipes in our mechanical room gives a hint about the detail available. If users can roam freely, they will no doubt move to areas where not enough detail exists to adequately fill the screen. In this scene, we can see the needle on the pressure gauge, but the numbers aren't sampled well enough to read. (The scene is rendered from the laser intensity image, courtesy of 3rdTech.)

sharing. May we acquire models of Disney World and share them free of charge? It's unlikely. Say you own a Lamborghini Miura—are you allowed to sell electronic models of it? How does that differ from selling copies of a book that you own?

The level of detail in acquisition is problematic. With detailed 3D models, viewers now have the opportunity to go wherever they please. Often a viewer will go up close to Jefferson's bookshelf and try to read the titles, but find that the data aren't at high enough resolution to do so. This happens over and over again, whether it's an attempt to look at the letter on Jefferson's table, see where the needle is pointing on the pressure gauges in mechanical rooms (see Figure 4), or look at the dashboard on the Mercedes in the crash scene. The data are limited in resolution, thus the viewer should be limited to viewing positions that don't push the data past its information conveyance point. To expand the possible viewing areas requires more detailed acquisition, a problem aided by advances in capture device resolution.

Three-dimensional scene acquisition does have some limitations, but its resolution, range, quality, and time required for capture are continuously improving. Modeling methods are improving in accuracy and speed (real-time range registration has been accomplished in conjunction with real-time range acquisition[11]). Rendering hardware has improved dramatically in speed and capability over the past few years. These improvements allow us to view highly detailed but nonetheless limited real-world 3D scenes. The color is accurate, the geometry is acceptable, and the realism is unmatched. With perseverance in hardware and software development, the views will get faster, more detailed, more realistic, and more compelling. ■

Acknowledgments

We're grateful to all of our collaborators, especially Leonard McMillan, Voicu Popescu, David McAllister, Chris McCue, Manuel Oliveira, Chun-Fa Chang, Paul Rademacher, and Qiong Han. Fellow faculty members Gary Bishop, Henry Fuchs, Nick England, David Luebke, John Eyles, and John Poulton provided many insights. Kurtis Keller provided excellent engineering support. We also thank John Thomas, our lab manager. Without him, none of our projects would succeed.

Financial support was provided by DARPA order number E278, the Department of Energy's Accelerated Strategic Computing Initiative program, NSF grant numbers MIP-9612643 and ACR-9876914, and the NSF Science and Technology Center for Graphics and Visualization.

References

1. M. Levoy et al., "The Digital Michelangelo Project: 3D Scanning of Large Statues," *Proc. Siggraph 2000*, ACM Press, New York, 2000.
2. D. Miyazaki et al., "The Great Buddha Project: Modeling Cultural Heritage through Observation," *Proc. 6th Int'l Conf. Virtual Systems and MultiMedia* (VSMM 2000), 2000, pp. 138-145.
3. F. Bernardini, J. Mittleman, and H. Rushmeier, "Case Study: Scanning Michelangelo's Florentine Pièta of the Cathedral," *Proc. Siggraph 98*, ACM Press, New York, 1998, p. 281. See also http://www.research.ibm.com/pieta.
4. M. Garland and P.S. Heckbert, "Surface Simplification Using Quadric Error Metrics," *Proc. Siggraph 97*, ACM Press, New York, 1997, pp. 209-216.
5. L. McMillan and G. Bishop, "Plenoptic Modeling: An Image-Based Rendering System," *Proc. Siggraph 95*, ACM Press, New York, 1995, pp. 39-46.
6. C.F. Chang, G. Bishop, and A. Lastra, "LDI Tree: A Hierarchical Representation for Image-Based Rendering," *Proc. Siggraph 99*, ACM Press, New York, 1999, pp. 291-298.
7. S. Rusinkiewicz and M. Levoy, "QSplat: A Multiresolution Point Rendering System for Large Meshes," *Proc. Siggraph 2000*, ACM Press, New York, 2000, pp. 343-352.
8. V. Popescu et al., "WarpEngine: An Architecture for the Post-Polygonal Age," *Proc. Siggraph 2000*, ACM Press, New York, 2000, pp. 433-442.
9. D.N. Wood et al., "Surface Light Fields for 3D Photography," *Proc. Siggraph 2000*, ACM Press, New York, 2000, pp. 287-296.
10. W.C. Chen, R. Grzeszczuk, and J.-Y. Bouguet, "Light Field Mapping: Hardware-Accelerated Visualization of Surface Light Fields, in Acquisition and Visualization of Surface Light Fields," *Proc. Siggraph 2001*, ACM Press, New York, 2001.
11. S. Rusinkiewicz and M. Levoy, "Efficient Variants of the ICP Algorithm," *Proc. Three-Dimensional Imaging and Modeling* (3DIM-2001), IEEE CS Press, Los Alamitos, Calif., 2001, pp. 145-152.

Readers may contact Nyland at CB 3175, Univ. of North Carolina, Chapel Hill, NC 27599, email nyland@cs.unc.edu.

Contact department editors Rhyne and Treinish by email at rhyne@siggraph.org and lloydt@us.ibm.com.