2006

Efficient Large Scale Acquisition of Building Interiors

Gleb Bahmutov

Report Number:
06-017
EFFICIENT LARGE SCALE ACQUISITION
OF BUILDING INTERIORS

Gleb Bahmutov
(Ph.D. Thesis)

Department of Computer Sciences
Purdue University
West Lafayette, IN 47907

CSD TR #06-017
August 2006
EFFICIENT LARGE SCALE
ACQUISITION OF BUILDING INTERIORS

A Thesis
Submitted to the Faculty
of
Purdue University
by
Gleb Bahmutov

In Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

December 2006
Purdue University
West Lafayette, Indiana
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF FIGURES</th>
<th>iii</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>viii</td>
</tr>
<tr>
<td>CHAPTER 1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1. Problem statement</td>
<td>1</td>
</tr>
<tr>
<td>1.2. Approach</td>
<td>3</td>
</tr>
<tr>
<td>1.3. Contributions</td>
<td>6</td>
</tr>
<tr>
<td>1.4. Organization</td>
<td>12</td>
</tr>
<tr>
<td>CHAPTER 2. PRIOR WORK</td>
<td>13</td>
</tr>
<tr>
<td>2.1. Acquired dense depth</td>
<td>13</td>
</tr>
<tr>
<td>2.2. User-specified coarse depth</td>
<td>14</td>
</tr>
<tr>
<td>2.3. No depth</td>
<td>15</td>
</tr>
<tr>
<td>2.4. Interactive modeling</td>
<td>16</td>
</tr>
<tr>
<td>CHAPTER 3. SYSTEM OVERVIEW</td>
<td>22</td>
</tr>
<tr>
<td>CHAPTER 4. MODELCAMERA ACQUISITION DEVICE</td>
<td>25</td>
</tr>
<tr>
<td>4.1. Design considerations</td>
<td>25</td>
</tr>
<tr>
<td>4.2. The ModeICamera</td>
<td>26</td>
</tr>
<tr>
<td>4.3. Calibration</td>
<td>38</td>
</tr>
<tr>
<td>4.4. Depth acquisition</td>
<td>43</td>
</tr>
<tr>
<td>CHAPTER 5. SECTION ACQUISITION: DEPTH ENHANCED PANORAMA</td>
<td>46</td>
</tr>
<tr>
<td>5.1. Frame registration</td>
<td>46</td>
</tr>
<tr>
<td>5.2. Frame merging</td>
<td>53</td>
</tr>
<tr>
<td>5.3. Real-time DEP visualization</td>
<td>56</td>
</tr>
<tr>
<td>CHAPTER 6. MODEL ASSEMBLY</td>
<td>61</td>
</tr>
<tr>
<td>6.1. Proxy fitting</td>
<td>61</td>
</tr>
<tr>
<td>6.2. Embedded detail</td>
<td>63</td>
</tr>
<tr>
<td>6.3. Section registration</td>
<td>66</td>
</tr>
<tr>
<td>CHAPTER 7. MULTI-VIEWPOINT MODELING</td>
<td>71</td>
</tr>
<tr>
<td>7.1. Handheld structured scene acquisition</td>
<td>73</td>
</tr>
<tr>
<td>7.2. Run-time multiple DEP merging</td>
<td>80</td>
</tr>
<tr>
<td>CHAPTER 8. RESULTS</td>
<td>88</td>
</tr>
<tr>
<td>CHAPTER 9. CONCLUSIONS AND FUTURE WORK</td>
<td>99</td>
</tr>
<tr>
<td>9.1. Future work</td>
<td>101</td>
</tr>
<tr>
<td>LIST OF REFERENCES</td>
<td>104</td>
</tr>
<tr>
<td>VITA</td>
<td>109</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1-1 ModelCamera acquisition device mounted in parallax free pan-tilt bracket. Shaft encoders (1, 2) report current tilt and pan angles. Laser diode (4) generates a beam that is split into an 11x11 matrix pattern using a custom diffraction grating (3).</td>
<td>7</td>
</tr>
<tr>
<td>Figure 1-2 Visualization of epipolar segments in a frame acquired with the ModelCamera aimed at a white wall, full frame (top) and magnified frame fragment (bottom). An epipolar segment is the locus of possible frame locations of the corresponding laser dot. The disjoint epipolar segments make depth extraction efficient (1-D search) and robust (unambiguous).</td>
<td>8</td>
</tr>
<tr>
<td>Figure 1-3 Acquisition of an indoor scene. The operator sweeps the scene by panning and tilting the acquisition device. The scene model is built from dense color and sparse depth frames at the rate of 5 frames per second.</td>
<td>9</td>
</tr>
<tr>
<td>Figure 1-4 Photograph of the operator feedback display. The evolving model is rendered continually to provide immediate feedback.</td>
<td>10</td>
</tr>
<tr>
<td>Figure 1-5 Model acquired with our system. A campus building (top left) was modeled by a two-person team in 40 hours. The model spans 6 floors with 20 attached rooms. Individual rooms (blue) and corridors (orange) are modeled with fitted proxy geometry enhanced with embedded detail.</td>
<td>11</td>
</tr>
<tr>
<td>Figure 2-1 A part of the “Jefferson’s Monticello” model [Williams 2003] is an example of a dense depth acquisition.</td>
<td>18</td>
</tr>
<tr>
<td>Figure 2-2 User-specified constraints and the computed 3D model using the ICARUS package [Hidalgo 2002].</td>
<td>19</td>
</tr>
<tr>
<td>Figure 2-3 Quicktime VR: individual photographs (top) are stitched into a complete color panorama (bottom) [Chen 1995].</td>
<td>20</td>
</tr>
<tr>
<td>Figure 2-4 Real-time 3D model acquisition system presented at SIGGRAPH 2002 [Rusinkiewicz et al 2002].</td>
<td>21</td>
</tr>
</tbody>
</table>
Figure 3-1 Section acquisition. A raw section model is acquired in 15 minutes at 5 fps with operator guidance........................................23

Figure 3-2 Model assembly. A raw section is refined and attached to the building model with operator input in 5 minutes. .........................24

Figure 4-1 Depth extraction by triangulation. The laser beam AB projects to the epipolar segment ab. Assuming that the laser beam intersects the scene surface at A, a bright laser dot is found at a. The point A (its depth za) is computed by intersecting the ray that originates at the video camera center of projection COP with the laser beam.............................................29

Figure 4-2 ModelCamera prototype 1. The video camera is surrounded by 16 laser diodes similar to the ones used in common laser pointers. ..........................................................30

Figure 4-3 Epipolar segment visualization for ModelCamera prototype 1 (full frame, top, and magnified frame fragment, bottom). The epipolar segments are drawn over a typical frame. A laser dot is detected along each epipolar segment.................................31

Figure 4-4 Visualization of the dot pattern obtained when the ModelCamera Prototype 1 is placed at three distances in front of a wall. For clarity only the 4 corner laser beams are drawn. .................................................................32

Figure 4-5 ModelCamera prototype 2 with the 16 individual laser diodes mounted using a rapid-prototyped plastic plate.........................33

Figure 4-6 ModelCamera prototype 3. A single compact laser diode produces a powerful beam split in a 7x7 pattern using a diffraction grating. Secondary dots are blocked by a wooden screen. .................................................................35

Figure 4-7 Visualization of epipolar segment pattern for prototype 3 (top full frame, bottom magnified frame fragment). The epipolar lines are clipped to segments using z = 50 and z = 350 cm planes. .................................................................36

Figure 4-8 ModelCamera Prototype 3 mounted on tripod using a parallax-free pan-tilt bracket. .................................................................37

Figure 4-9 Typical frame used for epipolar segment calibration. A brute force search algorithm finds the 11x11 brightest dots..................39

Figure 4-10 Epipolar segments fitted through each of the 11 x 11 groups of 200 dots.................................................................40
Figure 4-11 Laser beam 3D calibration. Setup with projected grid (left), frame used for dot detection (middle) and corresponding frame with projected grid for extrinsic calibration (right). ..........41

Figure 4-12 An intensity plot along a typical epipolar line with dot and false peaks. Line indicates minimum intensity threshold. .................45

Figure 5-1 DEP acquisition, visual feedback and resulting plant DEP. .................47

Figure 5-3 Registration pattern consisting of horizontal (red) and vertical (green) pixel segments using ModelCamera prototype 3. The frame was blurred and the dots were masked off (magenta). .....49

Figure 5-4 Plot of color error around the color registration solution as a function of pan and tilt angles measured in degrees...............50

Figure 5-5 Screen capture showing operator feedback showing the current frame (bottom left), the last registered frame (red rectangle), and depth samples (blue). .........................................................51

Figure 5-6 Room DEP acquired using prototype 3 with color registration. ......52

Figure 5-7 Panorama face without blending (top) and panorama face blended in real time (bottom). ........................................................55

Figure 5-8 DEP splatting (left). Magnified fragment with highlighted splat boundaries (right). .................................................................57

Figure 5-9 DEP shown as a connected triangle mesh (left). Magnified fragment with superimposed wireframe (right). .......................................................58

Figure 5-10 DEP triangulation. One face of a DEP with center O is shown. The depth samples P₀, P₁, and P₂ are connected in a triangle obtained by Delaunay triangulating their projections P₀', P₁', P₂' onto the face..................................................58

Figure 5-11 Depth samples are triangulated in 2D on the face of the panorama (orange) then the inferred connectivity is used to make the 3D mesh (blue). ........................................................60

Figure 6-1 Room section, and I, L, and T corridor sections. .........................62

Figure 6-2 Proxy depth sample pruning. Points close to the proxy planes (red) are discarded...............................................................63
<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 6-3</td>
<td>Corridor section with embedded detail. Representing the entire wall with a single plane make the water fountain appear flat (top left). The operator selects points (top right) which are kept in the section model to convey correct motion parallax (bottom).</td>
</tr>
<tr>
<td>Figure 6-4</td>
<td>Section registration: (top) original sections, (middle) sections after applying geometrical constraints, (bottom) connected sections.</td>
</tr>
<tr>
<td>Figure 6-5</td>
<td>Floor composed of 12 corridor and 4 room sections with the acquired point clouds</td>
</tr>
<tr>
<td>Figure 6-6</td>
<td>Wireframe visualization of proxy geometry of building model</td>
</tr>
<tr>
<td>Figure 7-1</td>
<td>Depth enhanced panorama of plant shown from viewpoint near (top) and far (bottom) from the acquisition viewpoint.</td>
</tr>
<tr>
<td>Figure 7-2</td>
<td>Handheld acquisition of structured scenes (top), and evolving set of depth images shown with green dots (bottom).</td>
</tr>
<tr>
<td>Figure 7-3</td>
<td>Surface fitting</td>
</tr>
<tr>
<td>Figure 7-4</td>
<td>Depth then color registration: pair of frames before depth registration (left), after depth registration (middle), and after color registration (right).</td>
</tr>
<tr>
<td>Figure 7-5</td>
<td>Depth asymmetric motions</td>
</tr>
<tr>
<td>Figure 7-6</td>
<td>Couch armrest modeled with depth images created on demand.</td>
</tr>
<tr>
<td>Figure 7-7</td>
<td>Model obtained from a pre-registered sequence of 2300 frames</td>
</tr>
<tr>
<td>Figure 7-8</td>
<td>Single DEP shown with splats highlighted in blue (top). Second DEP is acquired from a different position to fill in missing surfaces (bottom).</td>
</tr>
<tr>
<td>Figure 7-9</td>
<td>Input DEPs (top row), under sampling highlighted in red (bottom left), novel view rendered using sample selection based on sampling rate (bottom right).</td>
</tr>
<tr>
<td>Figure 7-10</td>
<td>Two input DEPs (red and blue) of a scene are combined to produce a novel view (green). The best sample for the desired view ray shown is the sample coming from the second DEP, although the sample coming from the first DEP is closer.</td>
</tr>
</tbody>
</table>
Figure 7-11 The desired view ray intersects undersampled triangles in all DEPs. We select the farthest sample, in this case the one coming from the second DEP. The farther sample invalidates the triangle coming from the first DEP, and indicates that the second DEP approximates the scene more closely. 86

Figure 7-12 Top row: two out of four input DEPs shown from their acquisition viewpoint. Middle row: each DEP shown from the desired view; the desired view is 151 cm and 91 cm away from the two acquisition points, respectively. Undersampling causes stretching. Bottom row: DEPs z-buffered together (left), and combined using GPU fragment shader which decides which sample to use based on both depth and sampling rate (right). 87

Figure 8-1 ModelCamera prototype 4 and the cart with PC can be easily moved to a new scanning location. 91

Figure 8-2 Captured corridor section rendered from a point close to the acquisition location (above) and from an extreme view (below) 92

Figure 8-3 Rendering of the basement floor from novel locations 93

Figure 8-4 Captured room section rendered from two novel viewpoints 94

Figure 8-5 Model of an office from two viewpoints 95

Figure 8-6 Room model rendered from two viewpoints with and without wireframe mesh 96

Figure 8-7 Basement floor model rendered from novel viewpoint 97

Figure 8-8 Corridor section with fragmented geometry 98
ABSTRACT

Bahmutov, Gleb. Ph.D., Purdue University, December, 2006. Efficient Large Scale Acquisition of Building Interiors. Major Professor: Voicu Popescu.

The present study describes a novel system for creating digital 3D models of building interiors. The system is efficient, which allows modeling indoor environments with tens of rooms and thousands of square meters of floor space in a few days. The models obtained enable photorealistic virtual walkthroughs at interactive rates and are suitable for computer graphics applications such as virtual training, cultural heritage preservation, real-estate marketing, and video gaming.

The system implements an automated modeling pipeline that runs at interactive rates. The operator sweeps the scene with a novel acquisition device that consists of a video camera and a laser system positioned on a tripod in the center of the scene. The laser casts an 11x11 matrix pattern of beams into the field of view of the video camera. The laser dots are detected in the video frame and converted to 3D points by triangulation. By construction, the laser beams project onto the video frame as disjoint line segments which makes dot detection efficient and robust. The dense color (720x480) and sparse depth (11x11) frames are registered and merged into a 3D model at the rate of 5 frames per second. The evolving model is rendered continually to provide immediate feedback to the operator. The operator aims the acquisition device to capture missing surfaces and adapts the depth sampling resolution to the local scene complexity. A model section (e.g. a room or a corridor segment) is created in minutes. The building
CHAPTER 1. INTRODUCTION

This thesis describes a novel system for efficient modeling of large-scale building interiors.

1.1. Problem statement

Geometry and color models of large scale indoor environments are invaluable to numerous applications in science, engineering, defense, art, and entertainment. Simulations of fire propagation in models of actual buildings lead to safer constructions. A model that captures the individual furniture configuration of hundreds of rooms enables a high-fidelity simulation of the propagation of an airborne contaminant. Virtual training of emergency response personnel is more effective in environments that capture the true complexity of real-world scenes. The ability of modeling large scale indoor environments enables virtual tourists to explore the buildings of an entire historic district. Real estate designers, constructors and marketers benefit from photorealistic interactive architectural walkthroughs. Insurance agents could use before and after models for estimating damages and validating claims.

All these applications require a detailed description of the color and geometry of the scene. The traditional approach of manual modeling using CAD or animation software requires a huge time investment as well as artistic talent. Even so, the resulting models fail to capture the full complexity of the real world. Automated modeling based on depth and color acquisition is a promising alternative that is being investigated in engineering, computer vision, and computer graphics.
model is assembled from sections by leveraging same-plane constraints specific to indoor architectural environments and minimal operator input.

We validated the system by modeling the interior of a large building on campus. The model spans 6 floors consisting of 20 rooms and corridors. The model was acquired by a two-person team with a single acquisition device in 40 hours. To the best of our knowledge, the model is the largest reconstruction of an indoor scene to date.
Good results have been obtained in “outside-looking-in” scenarios where a small scene is acquired from outside its bounding volume (e.g. scanning a statue for a virtual museum or a piston for reverse engineering). However, many modeling applications involve large scenes that need to be acquired from within (e.g. scanning the interior of a building for virtual training of emergency response personnel). Such “inside-looking-out” modeling is complicated by the sheer size of the scene, by the lack of control over acquisition conditions, and by the conundrum of occlusions. While digital photographs and video sequences suffice to show the scene in outside-looking-in scenarios, a 3D model of the scene is indispensable in the inside-looking-out case. For example an ancient mask can be captured well in minutes with 10 high-resolution digital photographs, but a 3D model is indispensable to allow virtual visits to historic homes.

The specific objectives of this work were:

1. **Inside-looking-out modeling.** The system should allow modeling the interior of multistory buildings that consist of rooms linked by corridors. Examples include office buildings, hospitals, hotels, and apartment buildings.

2. **Efficient large-scale modeling.** The system should allow modeling the interior of an entire building in days. For example, a 6-story 100-room building should be modeled in a week.

3. **Support for 3D computer graphics applications.** The acquired models should enable photorealistic virtual walkthroughs of the building interiors at interactive rates.
1.2. Approach

To achieve these objectives we have developed a system that implements an automated modeling pipeline which runs at interactive rates. The modeling pipeline is based on a novel acquisition device called the ModelCamera, see Figure 1-1.

The ModelCamera consists of a video camera and an attached laser system. The laser casts an 11x11 matrix pattern of beams into the field of view of the camera. The laser beams produce bright dots that uniformly sample about half of the video frame. The ModelCamera is connected to a host computer through a standard interface. The laser dots are detected in the video frame and converted into depth samples by triangulation between the laser beam and the video camera optical ray. By construction, the 50-350cm segment of each laser beam projects to a disjoint line segment in the video frame (Figure 1-2). We call the laser beam projection an epipolar segment, by analogy with stereo systems. The disjoint epipolar segments make dot detection efficient and robust. The ModelCamera robustly acquires dense color (720x480 pixels less the pixels covered by the dots) and sparse depth (11x11 depth samples) at video rate (15 Hz for our camera).

The operator sweeps the scene with the ModelCamera (Figure 1-3). We position the device on a sturdy tripod in the center of the scene using a parallax-free pan-tilt bracket. By this design, all frames share the same center of projection, allowing their registration using either color information or hardware devices attached to the bracket, similar to the registration of color-only panoramas. The frames are registered in a common coordinate system and integrated into a model at the rate of 5Hz. The evolving model is rendered continually to provide immediate operator feedback (Figure 1-4). The operator avoids over-scanning the simple parts of the scene (e.g. walls) and concentrates on the parts of the
scene with higher geometric complexity (e.g. bookshelves, plants). The building model is assembled from sections (e.g. rooms, corridor segments) which are aligned by leveraging same-plane constraints specific to indoor environments. An example of a model acquired with our system is shown in Figure 1-5. Since a tripod is used during the acquisition the quality of the model is the highest for scenes where most of the surfaces are visible from a single acquisition point.

The essence of our modeling approach is to replace slow and unreliable dense depth acquisition characteristic to prior systems with efficient and robust sparse depth acquisition. Sparse depth has great raw modeling power and allows efficiently modeling regions with simple geometry which constitute a significant fraction of typical indoor scenes.

1.2.1. Modeling power

A rough analysis shows that sequences of sparse depth frames have ample modeling power. In 15 minutes, the ModelCamera acquires 15mins x 60s x 5Hz x 11x11 dots x 80% dot detection rate = 435,600 raw depth samples. Relying on real-time feedback, the operator avoids over-sampling low curvature surfaces. This ensures that most of the depth samples are used in the refined geometric model. Assuming that only half of the raw depth samples are kept, and counting two triangles for each depth sample, the ModelCamera acquires a section model of 435,600 triangles in 15 + 5 = 20 minutes, which includes 5 minutes for post-processing the raw section model.

1.2.2. Structured vs. unstructured scenes

We distinguish between structured and unstructured scenes. Structured scenes consist of large surfaces such as doors, walls, and furniture. Unstructured scenes
contain many small surfaces, for example a plant, a messy bookshelf, or coats on a rack. The distinction is important for applications and facilitates automated modeling.

While a structured scene can be approximated well with a compact geometric model, achieving the same level of approximation for unstructured scenes requires a model of substantial complexity. Most graphical applications that involve large-scale inside-looking-out scenes do not need and cannot afford highly detailed models for the unstructured sections of the scene. For example, when the goal is rendering, as for example in virtual training of emergency response personnel, a complete model of every leaf of every plant would overwhelm the graphics engine and would interfere with the primary task of providing the trainee with images of the environment at interactive rates. Our modeling system supports fast approximate modeling of unstructured scenes, with the option of adding detail at proportional cost.

From the modeling perspective, it makes sense to take advantage of the relative simplicity of the structured sections of the scene, which represent a significant fraction of indoor scenes. Structured scenes do not need the same depth sampling resolution as unstructured scenes. A ModelCamera frame uniformly samples the scene geometry over a large solid angle (approximately half of the horizontal field of view and almost the entire vertical field of view). When the scene depth varies little, the dots are sufficient to obtain a quality surface approximation by interpolation.

Unlike prior automated modeling techniques, our interactive modeling paradigm gives the operator the information and flexibility to acquire and model the two types of scenes differently: unstructured scenes are acquired using ModelCamera on a tripod, relying on the same center of projection principle to register individual frames, and are modeled as a depth enhanced panorama;
structured scenes are acquired in handheld mode, relying on frame coherence and surface continuity to register frames, and are modeled using per-pixel depth images created on the fly.

This work concentrates on modeling of the unstructured scenes, the most difficult and general case of acquisition. We treat the acquisition of structured scenes as an example of the multiple viewpoint modeling designed to overcome the problem of modeling occluded parts in the scene.

1.3. Contributions

The interactive modeling pipeline is enabled by the following research and implementation contributions:

1. Robust dense color (~720x480) and sparse depth (11x11) at interactive rates (15fps).

2. Efficient wide solid angle (~20°x20°) light pattern with unambiguous detection guaranteed by construction leveraging epipolar constraints.

3. Efficient and effective scene section modeling using depth-enhanced panoramas (DEPs); a DEP is a texture-mapped triangle mesh acquired from a single viewpoint which supports incremental construction from uneven depth sampling and photorealistic rendering at interactive rates.

4. Complete modeling system with custom acquisition device and software for calibration, depth extraction, incremental modeling, visualization, and model assembly.
Figure 1-1 ModelCamera acquisition device mounted in parallax free pan-tilt bracket. Shaft encoders (1, 2) report current tilt and pan angles. Laser diode (4) generates a beam that is split into an 11x11 matrix pattern using a custom diffraction grating (3).
Figure 1-2 Visualization of epipolar segments in a frame acquired with the ModelCamera aimed at a white wall, full frame (top) and magnified frame fragment (bottom). An epipolar segment is the locus of possible frame locations of the corresponding laser dot. The disjoint epipolar segments make depth extraction efficient (1-D search) and robust (unambiguous).
Figure 1-3 Acquisition of an indoor scene. The operator sweeps the scene by panning and tilting the acquisition device. The scene model is built from dense color and sparse depth frames at the rate of 5 frames per second.
Figure 1-4 Photograph of the operator feedback display. The evolving model is rendered continually to provide immediate feedback.
Figure 1-5 Model acquired with our system. A campus building (top left) was modeled by a two-person team in 40 hours. The model spans 6 floors with 20 attached rooms. Individual rooms (blue) and corridors (orange) are modeled with fitted proxy geometry enhanced with embedded detail.
CHAPTER 2. PRIOR WORK

Modeling large scale indoor scenes requires capturing color and depth data from multiple viewpoints. The wide availability of photo and video cameras makes acquisition of high quality color an easy task. Sampling the scene geometry is far more challenging. This discussion of prior work groups automated modeling methods according to the technique employed for depth acquisition. We review methods based on acquiring dense depth maps, methods that rely on coarse geometry specified by the user, methods that omit depth acquisition altogether, and methods that model at interactive rates.

2.1. Acquired dense depth

Depth from stereo, triangulation laser rangefinding, and time-of-flight laser rangefinding technologies acquire dense, accurate depth maps that can be converted into high-quality models. Examples include digitizing Michelangelo’s statues [Levoy 2000, Bernardini 2002], Jefferson’s Monticello [Williams 2003], cultural treasures of Ancient Egypt [Farouk 2003], the Parthenon [Stumpfel 2003], and the ancient city of Sagalassos [Pollefeys 2001, 2002].

A disadvantage common to all modeling systems that acquire dense depth is the long per-view acquisition time, which limits the number of views. This in turn leads to incomplete models, especially in the inside-looking-out case where the device is surrounded by the scene (Figure 2-1). Another disadvantage is the high equipment cost.
1.4. Organization

The remainder of this document is organized as follows. Chapter 2 reviews prior scene modeling work. Chapter 3 describes the overview of the acquisition and modeling system. Chapter 4 describes the principles and successive generations of the acquisition device we have designed. Chapter 5 describes modeling of an individual scene using the Depth Enhanced Panorama (DEP) data structure. Chapter 6 describes post-processing of individual DEPS and assembling the complete building model. Chapter 7 discusses two approaches we have investigated for modeling scenes from multiple viewpoints. Chapter 8 describes the modeling of a large building we have carried out to validate the proposed system. Finally, in chapter 9 we state the conclusions learnt from this work and possible directions for future work.
The goal of our work has not been to devise new depth acquisition technology, but rather to achieve inside-looking-out modeling at interactive rates. No dense depth acquisition device is sufficiently fast, compact, robust, and inexpensive. Should a device emerge that satisfies the needs of inside-looking-out interactive scene modeling, it will be integrated with our system.

The problem of depth acquisition continues to be investigated from several directions. Computer vision researchers continually increase the quality and robustness and decrease the acquisition time of depth maps extracted from images, as recently reported in [Darabiha 2003], [Yang 2003], [Zhang 2003], and [Davis 2003]. The Zcam technology [3dvsystems] acquires depth in parallel over the entire view frustum. The Zcam is used as an add-on to studio cameras for real time depth keying [Gvili 2003]. Its depth resolution is too low for modeling. Helmholtz stereopsis [Zickler 2002] is a new depth extraction technique that provides normal and depth estimates—thus combining the advantages of classic and photometric stereo—without restricting the surface reflection model. The problem of quickly finding correspondences remains.

2.2. User-specified coarse depth

Another solution to the depth acquisition problem is manual geometry-data entry. An example is the Façade architectural modeling system in which the user creates a coarse geometric model of the scene that is texture mapped with photographs [Debevec 1996]. The geometric part of the hybrid geometry-image-based representation is created from user input in [Hidalgo 2002] (Figure 2-2). In view morphing [Seitz 1996], the user specifies depth in the form of correspondences between reference images.

Another example is image-based editing [Anjyo 1997, Oh 2001], which builds 3D models by segmenting images into sprites that are mapped to separate planes.
User-specified coarse depth systems take advantage of the user's knowledge of the scene, which allows him to maximize the 3D effect while minimizing the amount of depth data. The disadvantage of the approach is the need for manual input, severely limiting the complexity of the model. An example 3D model derived from user-specified geometric constraints is shown in Figure 2-2.

2.3. No depth

Some modeling techniques bypass depth acquisition altogether. QuickTime VR panoramas [Chen 1995] are 2D ray databases that store a dense sampling of the rays passing through one point. They are constructed by stitching together same-center-of-projection images (Figure 2-3). They support viewing the scene from this point in any desired direction. Panoramas have the advantages of rapid, inexpensive acquisition and of interactive photorealistic rendering, which makes them the only inside-looking-out modeling technique widely used. Panoramic image mosaics [Szeliski 1996, Zoghiami 1997, Coorg 1998] relax the same-center-of-projection constraint and stitch seamlessly hundreds of images that were taken from the approximately same location. Omnidirectional imaging enables the direct acquisition of wide-field-of-view 2D data base of concurrent or quasi-concurrent rays using dioptric (e.g. fish eye lens) or catadioptric cameras (e.g. parabolic mirror, planar pinhole camera cluster) [Nayar 1997, Kropp 2000, Aliaga 2001, Geyer 2002]. The disadvantage of the 2D ray database approach is the lack of geometry. This precludes physical simulation applications, and severely restricts computer graphics applications by not supporting view translations and depriving the user of motion parallax, an important cue in 3D scene exploration.

Light fields [Levoy 1996, Gortler 1996] are 4D ray databases that allow a scene to be viewed from anywhere in the ray space. An advantage of light field rendering is support for view dependent effects, such as reflection and refraction.
Light fields are constructed from a large set of registered photographs. Acquiring and registering the photographs is challenging. Another disadvantage is that the database is impractically large for complex scenes. Like panoramas, lightfields do not support applications that require geometry.

2.4. Interactive modeling

If a small part of the scene is acquired at each view, the per-view depth acquisition task is simplified and can be carried out by portable devices. Several hand-held depth acquisition devices have been developed. One architecture is a fixed camera and a mobile light-pattern source. One variant [Takatsuka 1999] uses a hand-held laser point projector on which three green LED's are mounted. The position of the LED's in the camera frame is used to infer the position and orientation of the laser beam. The red laser dot is detected in the frame and then triangulated as the intersection between the pixel ray and the laser beam. Another variant [Bouguet 1999] extracts depth from the shadow of a rod captured by a camera under calibrated lighting. Another architecture [Borghese 1998] uses two cameras mounted on a tripod and a hand-held laser point projector. The main problem with these systems is that they are limited to a single view by the fixed camera.

Some systems obtain frame registration from external trackers (e.g. electromagnetic senders and receivers, or mechanical arms) and concentrate on integrating the depth and color data of each frame into a texture mapped geometric model [Fisher 1996], [Hilton 2000]. Hebert [2001] proposes a system where the operator can freely change the view. The device consists of two cameras and a cross-hair laser light projector. Frame to frame registration is achieved using a set of fixed points projected with an additional, fixed laser system. The fixed points are easy to discern from the cross-hair and act as fiducials. The system is not well suited for large scenes, since a large number of
fiducials would be needed. It acquires depth only over a very narrow field of view at each frame, which implies long acquisition times in the case of complex scenes. It does not acquire color.

Rusinkiewicz et al. [2002] present an object modeling system based on structured light (Figure 2-4). The object is maneuvered in the fields of view of a fixed projector and camera. The frames are registered in real time using an iterative closest point algorithm. The evolving model is constructed in real time and is rendered to provide immediate feedback to the operator. The system is limited to the outside-looking-in modeling case and does not acquire color. A similar system is proposed by Koninckx [2003] where moving or deformable objects are captured in real time. The system acquires depth using a pattern of equidistant black and white stripes and a few transversal color stripes for decoding. The disadvantages of their system are limited acquisition range due to the fixed camera and projector configuration and the need for strict lighting control. Despite their shortcomings, both systems demonstrate the advantages of interactive modeling.
Figure 2-1 A part of the "Jefferson's Monticello" model [Williams 2003] is an example of a dense depth acquisition.
Figure 2-2 User-specified constraints and the computed 3D model using the ICARUS package [Hidalgo 2002].
Figure 2-3 Quicktime VR: individual photographs (top) are stitched into a complete color panorama (bottom) [Chen 1995].
Figure 2-4 Real-time 3D model acquisition system presented at SIGGRAPH 2002 [Rusinkiewicz et al 2002].
CHAPTER 3. SYSTEM OVERVIEW

Our modeling system targets buildings where each floor has a corridor with attached rooms. Examples include office buildings (the model shown in Figure 1-5 is a building that houses a department on our university campus), hospitals, hotels, and apartment buildings. Our system does not have the depth acquisition range needed for large indoor spaces such as theaters or warehouses. The corridors are assumed to have a rectangular cross-section. We handle corridor turns, loops, and junctions, as well as occasional objects or corridor sections with high geometric complexity.

The operator models a section of the building at a time. A section is a room or a corridor segment. A room is acquired from its center. Corridors are acquired by placing the ModelCamera acquisition device (Chapter 4) at corridor turns and junctions. Long corridor spans are split up in segments of approximately 8m. For each section, color and depth data is first acquired at interactive rates to create a raw model (Figure 3-1). Then the raw section model is post-processed and integrated into the building model (Figure 3-2).

Section acquisition (Chapter 5)

The ModelCamera is mounted on a tripod using a pan-tilt bracket. The bracket allows panning and tilting about the center of projection of the video camera. The operator pans and tilts the ModelCamera to sweep the scene. The same center of projection dense color and sparse depth frames are first registered in a common coordinate system using color data for prototype 3 and shaft encoder data for prototype 4. The registered depth and color frames are merged into an evolving model, which is displayed continually for immediate operator feedback.
The average frame processing time is 200ms. Relying on the instant feedback the operator samples each surface adequately.

*Model assembly (Chapter 6)*

A raw section model is processed by fitting a proxy to the depth samples. In the case of a room section the proxy is equivalent to an extruded floor plan. For corridors there are three types of section proxies: I, L, and T section. The unnecessary depth samples on the planes of the proxy are eliminated. However, geometric detail is preserved with triangular meshes embedded into the proxy. The final section models are added to the floor model, and the floor models are then stacked up to form the building model. Corridor sections are aligned and connected by leveraging same-plane constraints and with minimal operator input. Color is blended to alleviate color differences between the two sections spliced together. Room sections are slid into place such that the two faces of the door align. Post-processing a raw section model takes approximately 5 minutes.

![Diagram](image)

Figure 3-1 Section acquisition. A raw section model is acquired in 15 minutes at 5 fps with operator guidance.
Figure 3-2 Model assembly. A raw section is refined and attached to the building model with operator input in 5 minutes.
CHAPTER 4. MODELCAMERA ACQUISITION DEVICE

This chapter describes the design considerations that led to the development of the ModelCamera acquisition device, the earlier and current ModelCamera prototypes, the calibration procedures for the ModelCamera acquisition device, and the depth extraction procedure for the ModelCamera device.

4.1. Design considerations

The acquisition device was designed to satisfy the following requirements:

1. The acquisition device should capture both color and geometry from the same viewpoint. This avoids the difficult additional step of color to depth registration which plagues systems that acquire depth and color in a two passes using two sensors. First, color to depth registration is laborious because it implies manually establishing correspondences between the depth and color data. Second, the two sensors cannot be placed in exactly the same location and the residual offset between the two acquisition viewpoints prevents finding a perfect solution to the color to depth registration problem.

2. The acquisition device should be maneuverable and the modeling pipeline should run at interactive rates. This allows the operator to freely aim the acquisition device to target the desired part of the scene, for as little or as long as needed. The model is built incrementally and is available for inspection in real time. The operator can intuitively adapt the sampling resolution on the fly in order to maximize the impact of the acquired depth data and to optimize modeling efficiency. This is a radical departure from...
conventional sequential scanning which is inefficient since all surfaces are scanned at the same resolution regardless of the geometric complexity.

3. The acquisition device should sample the scene depth, albeit sparsely, over a large solid angle with every frame. This allows deriving a tight approximation of a structured scene patch (e.g. wall, floor, ceiling, or couch) from a single frame and allows quickly approximating a portion of an unstructured scene from a few frames, with the option of adding detail at linear time cost. The ensuing modeling efficiency is far greater than for example in the case of a conventional structured light scanner that employs a single plane of light and requires the operator to slowly sweep even simple surfaces such as walls.

4. The acquisition device should operate robustly under normal indoor conditions without requiring scene modifications such as controlling the lighting, moving objects around, or adding fiducials, modifications that greatly increase modeling time and prevent scaling to large indoor environments.

4.2. The ModelCamera

To meet the above requirements we have opted for a structured light approach to depth extraction. Compared to depth from stereo, structured light has the advantage of efficiency and robustness. Pairing a camera with a source of controlled light greatly simplifies the problem of searching for correspondences and allows extracting depth even in the absence of distinct scene features.

In order to meet the first requirement—single viewpoint depth and color acquisition—a structured light acquisition device has to acquire color (scene imaging) and depth (structured light pattern imaging) with the same camera, from the same viewpoint. The two options are acquiring color and depth simultaneously with each frame, and time multiplexing depth and color frames coupled with restricting the acquisition device motion to rotations about the center of projection of the camera. The first option has the disadvantage that the
light pattern has to share the frame with the scene color samples, but has the advantage that each frame acquires intrinsically registered depth and color data and therefore the acquisition device can be translated between frames.

In order to meet the maneuverability requirement the acquisition device should be compact and should acquire depth and color data as the operator aims it freely at the scene. This requirement together with the requirement that the acquisition device operates under the original indoor lighting conditions precludes the use of fragile and dim projectors as a source of structured light. Laser diodes have the advantage of compactness, sturdiness, and brightness.

For the automated modeling pipeline to run at interactive rates, each stage—depth and color acquisition, registration, modeling, and visualization—has to be efficient. Depth acquisition is the most important bottleneck, and, for the foreseeable future, the only approach for accelerating the pipeline to interactive rates is to decrease the amount of depth acquired with each frame.

The ModelCamera consists of a digital video camera and an attached laser system. The laser casts several beams into the field of view of the camera. The camera acquires frames of scene color enhanced with bright dots where a laser beam intersects a scene surface. The bright dots are the structured light pattern. The bright dots cover only a small subset of the frame pixels, so sufficient pixels remain for scene color acquisition.

Since the laser system is attached to the video camera, the projection of each laser beam onto the video frame remains constant as the operator moves the acquisition device. This means that each dot is known to move on a predetermined line in the video frame, which we call an epipolar segment, term borrowed from the geometry of two-camera systems used in depth from stereo. Each dot is found quickly and robustly with a 1D search along the epipolar
segment. Once the dot is located in the frame, it is converted to a 3D point (depth sample) by intersecting the video camera ray and the laser beam (Figure 4-1). Using several laser beams as opposed to a single beam or plane of light complicates the laser source but satisfies the requirement that depth be acquired over a large solid angle, which translates in modeling efficiency.

We have built 4 ModelCamera prototypes. All prototypes operate according to the principles described above. They differ by the number and configuration of the laser beams. The first two prototypes are now obsolete, but we briefly describe them here for completeness.

4.2.1. ModelCamera Prototype 1

The first ModelCamera prototype was designed to experiment with various configurations of the laser beams (Figure 4-2). A high-end consumer-level digital video camera (progressive scan, 3 CCDs, 720x480 resolution, FireWire interface, $1,500) is surrounded by 16 laser diodes. The diodes emit red laser light (wavelength of 635nm), have an emitting power of 5mW (class IIIa), have a spot size of 6mm/12mm at 5m/15m, and cost $15 apiece. The diodes are mounted in a matrix pattern around the camera using a wooden board and metal rods.

Each diode is held in place with a bracket that allows aiming the each laser beam independently. A favorable laser beam configuration is shown in Figure 4-3. The laser beams converge first to reach a highest depth sampling density and then diverge again. No two laser beams intersect, but epipolar segments do intersect, which occasionally leads to dot detection ambiguity. Figure 4-4 simulates placing the ModelCamera at three different depths in front of a wall and illustrates the beam pattern. The convergent and then divergent configuration of the lasers allows the operator to vary the sampling rate by moving the ModelCamera closer and farther from the surface to be acquired. A quick scan
from afar suffices for simple surfaces, such as walls, whereas a close, dense scan is needed for complex shapes.

The first ModelCamera prototype was used in freehand mode to model structured scenes, like the couch seen in Figure 4-3. Freehand modeling of structured scenes is described in Section 0. An important shortcoming of the first ModelCamera prototype is the lack of rigidity of the laser diode mount, which caused losing calibration frequently.

Figure 4-1 Depth extraction by triangulation. The laser beam AB projects to the epipolar segment ab. Assuming that the laser beam intersects the scene surface at A, a bright laser dot is found at a. The point A (its depth \( z_a \)) is computed by intersecting the ray that originates at the video camera center of projection COP with the laser beam.
Figure 4-2 Model Camera prototype 1. The video camera is surrounded by 16 laser diodes similar to the ones used in common laser pointers.
Figure 4-3 Epipolar segment visualization for ModelCamera prototype 1 (full frame, top, and magnified frame fragment, bottom). The epipolar segments are drawn over a typical frame. A laser dot is detected along each epipolar segment.
4.2.2. ModelCamera Prototype 2

For the second prototype the wooden board and metal rods were replaced with an L shaped Aluminum bracket that holds a rapid-prototyped plastic plate in which the diodes are screwed in (see Figure 4-5). The changes made the acquisition device more rigid and more lightweight. The lasers were all moved to one side of the camera, changing the epipolar segment configuration. The epipolar configuration is similar to the configurations for prototypes 3 and 4 and is discussed below, part of the description of those later prototypes.
Although the mechanical properties of the second prototype are greatly improved, the prototype performance suffers from the low quality of the laser diodes which produce an uneven, off-axis beam.

Figure 4-5 ModelCamera prototype 2 with the 16 individual laser diodes mounted using a rapid-prototyped plastic plate.

4.2.3. ModelCamera Prototype 3

Prototypes 3 and 4 are the prototypes that we currently use. Prototype 3 brings the substantial improvement of using a single, high-quality laser diode [Stockeryale] (see Figure 4-6) which replaces the individual, low-quality diodes of earlier prototypes. The beam is split into 7x7 evenly spaced beams using a diffraction grating. The diode with the diffraction grating weigh 100g and produce 49 1mW eye safe dots (class IIb). The drawbacks of the diffraction grating were increased red ambient light between dots, secondary dots, and lower individual
dot brightness. The secondary dots are blocked off using a wooden screen. The lower dot brightness level did not significantly change the dot detection success, causing difficulty only on the most optically absorbent scenes (such as green plants).

The epipolar segment configuration is shown in Figure 4-7. As the ModelCamera moves closer to the scene, the laser dots slide from the left to the right endpoint of each epipolar segment. The epipolar segments are disjoint, which eliminates the possibility of dot confusion.

The ModelCamera prototype 3 acquires frames of ~720x480 color and 7x7 depth samples at 15fps. The third prototype is used in handheld mode to model structured scenes (Figure 4-6) and on a tripod to model unstructured scenes (Figure 4-8). In order to facilitate registration and modeling of unstructured scenes, the ModelCamera motion is restricted to parallax-free pan and tilt rotations about the center of projection of the video camera using a special bracket. The bracket was adapted from a commercial tripod head for telephoto lenses [Wimberley] (see Figure 4-8). Frames are registered together the same way overlapping photographs are stitched together to form color panoramas [Chen 1995], by searching for the pan and tilt angles that minimize the color difference at the region of overlap.

Prototype 3 is light weight and inexpensive (the laser system costs $1,500). However, it has two important shortcomings. First, registration can only succeed for scenes with color variation (texture). Large spans of white wall pose a challenge. Second, the interbeam angle—which had to be chosen according to the available diffraction gratings—does not allow using the entire vertical field of view of the video camera (Figure 4-7).
Figure 4-6 ModelCamera prototype 3. A single compact laser diode produces a powerful beam split in a 7x7 pattern using a diffraction grating. Secondary dots are blocked by a wooden screen.
Figure 4-7 Visualization of epipolar segment pattern for prototype 3 (top full frame, bottom magnified frame fragment). The epipolar lines are clipped to segments using $z = 50$ and $z = 350$ cm planes.
4.2.4. ModelCamera Prototype 4

We built a fourth ModelCamera prototype that overcomes these problems (Figure 1-1). First, the diffraction grating was designed to optimize the use of the video frame. The interbeam angles were computed to use the entire vertical field of view of the camera (Figure 1-2). A powerful laser generates $11 \times 11$ 5mW dots, which is the upper limit for the power of eye safe visible lasers (class IIIa). Each dot is bright enough to be detected even on highly absorbent surfaces. Second, in order to be able to register frames robustly even in the absence of color, we enhanced the parallax free bracket with shaft encoders, which report the current pan and tilt angles (Figure 1-1). These modifications come at a substantial increase in cost, the ModelCamera prototype 4 costs $20,000, not including the one time tooling charge for the custom diffraction grating.
4.3. Calibration

Before the ModelCamera can be used to acquire frames of dense color and sparse depth one needs to know the equation of the laser beams into the coordinate system of the video camera which enables triangulation. The ModelCamera is calibrated in 3 steps, followed by another 3 steps for calibrating the parallax-free bracket.

4.3.1. Optical calibration

The video camera intrinsic parameters, including distortion coefficients, are found using standard camera calibration [Bouguet]; the calibration error is 0.075 pixels, computed as the average image plane distance between the projection of a 3D point (checkerboard corner) and the location where it was found in the image.

The laser beam equations in the local video camera coordinate system are found in two steps.

4.3.2. Epipolar segment calibration

Since the laser beams are fixed with respect to the camera, their image plane projection is confined to a fixed epipolar line. The second calibration step finds the epipolar line of each laser beam. 200 frames are recorded by aiming the camera at a white wall from various distances. The 11x11 laser dots are found in each undistorted frame (Figure 4-9) with a brute force search for bright spots. The brute force search is aided by turning the ambient light off. A 2D line is least-squares fitted through each group of 200 points, with an average error of 0.2 pixels (Figure 4-10).
Figure 4-9 Typical frame used for epipolar segment calibration. A brute force search algorithm finds the 11x11 brightest dots.
Figure 4-10 Epipolar segments fitted through each of the $11 \times 11$ groups of 200 dots.
4.3.3. Laser beam 3D calibration

The epipolar segment together with the video camera COP define an epipolar plane which is known to contain the laser beam. In the third calibration step the line equation of each laser beam is finalized. For this the ModelCamera is placed on a tripod and is aimed at a white wall (Figure 4-11). The laser dots are found on their respective epipolar lines. Then the laser is turned off and a checkerboard is projected orthographically onto the wall. Using the checkerboard and the known intrinsics the camera pose is found in the coordinate system defined by the checkerboard. A 3D point is found for each laser beam by intersecting the checkerboard plane xy with the camera ray where the dot was found. The tripod is moved to other locations to collect several (~10) 3D points for each laser beam. A line is least squares fitted to the points of each laser beam with an error of 2.5mm, computed as the average distance from the points to the line.

![Laser beam 3D calibration. Setup with projected grid (left), frame used for dot detection (middle) and corresponding frame with projected grid for extrinsic calibration (right).](image)

4.3.4. Parallax free positioning

The first of three steps that calibrate the parallax-free bracket positions the video camera such that the pan and tilt axes of the bracket pass through its center of projection. The camera is positioned by hand by trial and error: we slide the
camera forward-backward and up-down on the bracket until panning and tilting do not show motion parallax between a near (30cm) and a far (350cm) object. We estimate that the accuracy achieved is better than 5mm since a translation of 5mm away from the calibrated position in either direction introduces clearly noticeable motion parallax when panning or tilting. This figure is confirmed by another indirect measurement. When scanning a white wall by panning the ModelCamera, the distance to the wall measured using the depth samples stays within 3mm.

4.3.5. Pan / tilt axes calibration

Once the ModelCamera is placed at the parallax free location we determine the equations for the two axes in the coordinate system of the video camera. The axes are found one at a time, with a similar procedure. As the camera is rotated around the targeted axis, several overlapping frames are registered by minimizing color error over three rotational degrees of freedom. The found angles are converted to a single rotation which gives the axis. The tilt axis is the same from one sequence of frames to the next. However, the pan axis depends on the initial tilt angle of the ModelCamera. In order to avoid recalibrating the pan axis, each sequence starts with the same tilt angle marked on the shaft encoder drum.

These five steps take approximately 20 minutes, and their result is reused in many acquisition sequences. Prototype 4 requires a sixth step that registers the frame time stamps (camera clock) with the shaft encoder time stamps (PC clock), and has to be performed for each acquisition sequence.

4.3.6. Shaft encoder delay calibration
The time registration is performed by finding the constant offset between the PC clock used to poll the shaft encoders and the camera clock used to time stamp the video frames. This allows synchronizing the angle readings with the frames without using an explicit (hardware) sync between the computer and the camera. The offset is determined by taking advantage of the fact that the shaft encoders can be polled very frequently—10 times each millisecond—which for our application is equivalent to instantaneous angle reads. The second fact used in the calibration of the clock offset is that video frames are acquired at evenly spaced time intervals.

The operator pans the camera over a part of the scene with high color detail which allows precise color registration. The shaft encoder angles are read in frequently (every millisecond) and stored in a buffer together with the PC time when they were acquired. Using the known pan axis, the frames are registered in 1D using color. Once the pan angle of a frame with camera timestamp $c$ is known, the angle is looked up in the buffer of shaft encoder angles. The corresponding PC clock timestamp $p$ is used to compute the delay as $p - c$. The precision of the calibration increases with panning speed, since this shortens the time interval where the angle values stay constant from a frame to the next. By panning 30 degrees in 4-5 seconds, we typically obtain 6 – 10 delay values agreeing within 1 millisecond. The calibrated delay is used during acquisition to look up the pan and tilt angles for each frame in a buffer indexed this time by the video frame timestamp.

4.4. Depth acquisition

Once the ModelCamera is calibrated, each point on each epipolar segment corresponds to a 3D point. The diffraction grating and the orientation of the laser head ensure that the epipolar lines are disjoint and as far apart from each other as possible. This avoids dot confusion and makes depth acquisition robust. Each
dot is found along its epipolar segment by searching for an intensity peak. The dot detector finds intensity peaks along the epipolar segments. In Figure 4-12 four peaks are found. The peak for the actual dot is located by examining the shape of the bright spot in 2D and eliminating peaks that do not pass symmetry tests.

Unstructured scenes make dot detection more challenging since their color can have patterns similar to the dots. In order to reduce the number of false positive laser dot detections, candidate dots have to pass two tests. Firstly, a new dot has to be within epsilon pixels of the location where it was found in the previous k frames. A legitimate jump from one surface to another is validated after k frames. Secondly, dots cannot be located at the same location in the cube map as the camera is rotated. This indicates confusing the dot with a scene feature. The number of false positives that pass these tests is negligible (in most DEPs the number was zero and less than 100 per DEP in others). Moreover, the false positives generate points that are clustered in front or behind the scene, which makes selecting and deleting them straightforward.

The average dot detection success rate is between 60% and 99% depending on distance and surface properties. Dot detection takes less than 5ms per frame (all timing information reported in this paper is for a 2GHz 2GB Pentium Xeon PC). The depth accuracy is a function of the dot detection accuracy, of the camera field of view, of the frame resolution, and of the baseline. For a baseline of 15 cm, a one-pixel dot detection error translates into a depth error of 0.1 cm at 50 cm, 0.35 cm at 100 cm, 1.5 cm at 200 cm and 3.5 cm at 300 cm. We estimated dot detection accuracy by scanning a white wall from several distances and measuring the out-of-plane displacements of the triangulated 3D points. At 200 cm, the average/maximum displacements were 0.33 cm/1.1 cm, which indicates a dot detection error of 0.5 pixels. Better results were obtained at shorter distances.
The calibrated ModelCamera acquires video frames enhanced with 7x7 or 11x11 depth samples. The operator can freely maneuver the device which allows sampling each surface at an appropriate resolution. The dense color and sparse depth frames are integrated into an evolving model in real time as follows.

Figure 4-12 An intensity plot along a typical epipolar line with dot and false peaks. Line indicates minimum intensity threshold.
CHAPTER 5. SECTION ACQUISITION: DEPTH ENHANCED PANORAMA

In order to support the interactive rate modeling pipeline we developed a method for incremental modeling that is efficient, accommodates uneven depth sampling characteristic to our efficient data acquisition, allows adding detail at linear cost, and allows visualizing the model in real time. Our inspiration came from color panoramas [Chen 1995], which are a powerful approach for modeling inside-looking-out scenes efficiently. We developed a Depth Enhanced Panorama (DEP) which is a single viewpoint texture mapped triangle mesh constructed incrementally by registering and merging dense color and sparse depth frames at interactive rates (Figure 5-1). DEPs maintain the advantages of color panoramas such as efficient modeling of large scale scenes. However, unlike regular color panoramas, DEPs allow viewpoint translation and store geometry explicitly which enables quantitative applications such as physical simulation.

5.1. Frame registration

Registration transforms the current frame data (sparse depth and dense color) from camera coordinates to world coordinates. Since the frames share a common center of projection, they can be registered using rotation angles around pan and tilt axes; the color from individual frames can be stitched into a consistent color panorama (Figure 5-2). We have developed two methods to compute the unknown angles: color registration using the texture observed in each frame (prototype 3) and hardware registration (prototype 4) using the shaft encoders. In both cases a frame is rejected if the camera has not moved between the frames by a minimum angle to avoid processing of redundant data.
Figure 5-1 DEP acquisition, visual feedback and resulting plant DEP.
Figure 5-2 Frame (red) registered to DEP cube map (blue) by finding pan and
tilt angles.

5.1.1. Color registration

We have developed a fast registration algorithm that minimizes a color error
function whose arguments are the pan and tilt angles. The error of a pixel in the
current frame is the Euclidian distance between its color and the color where it
projects in the cube map. Even small camera motions produce rapid, erratic
changes in color error. We reduce the variability and the noise by convolving the
frames with an 11x11 raised cosine filter.

We then select a registration pixel pattern in the current frame (Figure 5-3). The
pattern consists of horizontal and vertical segments that exhibit considerable
color variation. The pixels of a segment share the same row or column and thus
can be projected onto the cube map faces with an amortized cost of 3 additions
and 2 divisions. We minimize the sum of the square of the pixel errors by the
downhill simplex method. The dot pixels are excluded because their color comes
from the lasers, rather than from the scene. The simplex method does not require
derivatives, which are expensive to compute. The rotation angles from the
previously registered frame are used as an initial guess, in most cases the
search valley is smooth and the simplex quickly converges to the minimum solution (Figure 5-4).

Figure 5-3 Registration pattern consisting of horizontal (red) and vertical (green) pixel segments using ModelCamera prototype 3. The frame was blurred and the dots were masked off (magenta).

Color registration takes 150ms per frame and merging takes 50ms per frame, so the modeling rate is 5 frames per second. When there is sufficient color variation in the scene, the segments and the cube map faces are down sampled by a factor of up to 10, which accelerates DEP construction to 10 frames per second. The registration algorithm fails once in 100-300 frames on average. The operator easily regains registration by aligning the camera view with the last registered frame (Figure 5-5). An example of a captured DEP is shown in Figure 5-6.
Figure 5-4 Plot of color error around the color registration solution as a function of pan and tilt angles measured in degrees.
Figure 5-5 Screen capture showing operator feedback showing the current frame (bottom left), the last registered frame (red rectangle), and depth samples (blue).
Figure 5-6 Room DEP acquired using prototype 3 with color registration.
5.1.2. Hardware registration

The hardware registration allows successful acquisition of featureless scenes, common inside buildings. Each new frame is registered using the angles provided by the shaft encoders. The angles are looked up in the buffer using the frame timestamp and the calibrated shaft encoder delay. The angle lookup avoids the registration time cost and increases the acquisition speed to 15 frames per second, the progressive scan frame rate limit of the video camera. A complete cube map is acquired in 4-5 minutes.

5.2. Frame merging

Once the frame is registered its color and depth data is added to the model. The depth data is new, except for the rare occurrences when two dots sample the depth of the scene along the same direction, so the depth samples can be simply accumulated. The color data on the other hand is highly redundant with the color acquired by the previous frames. Instead of keeping the individual frames we merge the color data into a panorama, implemented as a cube map.

For efficiency we divide the panorama into square tiles. In a first step we determine the set of tiles affected by the current frame. In a second step the tiles in the set are filled in using the current frame. Modern consumer-level video cameras capture high quality color, but overlapping frames can have a considerably different brightness, as the camera automatically adjusts exposure. One option would be to disable the automatic adjustments, but this will lower the quality of the frames. We have opted for unevenly exposed but smoothly varying color data, which we achieve by blending overlapping frames on the fly (Figure 5-7).
A tile is filled with the color from a given video frame if it is empty, or if the frame’s brightness is higher than the tile’s current brightness. This approach works well indoors: detail is captured in the darker parts of the scene, while only saturating the fluorescent lights on the ceiling. Each tile is larger than its contribution to the panorama to allow efficient blending of tiles with its neighbors. We found that 32 by 32 pixel tiles, with an additional border of 16 pixels is a good compromise between processing speed and quality of the resulting color.

The bracket does not allow capturing color right above the camera. As a temporary solution we fill this gap with color from the surrounding regions. The tripod interferes with color acquisition directly beneath the camera. The hole in the floor can be filled in for corridors due to the repetitive nature of the texture, but filling in color for complex rooms is more challenging.
Figure 5-7 Panorama face without blending (top) and panorama face blended in real time (bottom).
5.3. **Real-time DEP visualization**

In our scanning system the operator can directly control the acquisition by observing the evolving model in real time. To provide immediate feedback to the operator during the acquisition we have developed two different visualization modes.

5.3.1. **Disconnected representation**

A disconnected visualization method for DEPs is similar to the splatting techniques of point-based modeling and rendering: QSplats [Rusinkiewicz '00], surfels [Pfister '00], and forward rasterization [Popescu '00]. None of these methods applies, since DEPs are sparsely populated with depth samples. Instead, we generate a texture-mapped square splat for each depth sample. The splat's size and normal are derived from the neighboring depth samples, located quickly using the quadtree data structure. The neighbors are triangulated and the normals of the triangles are averaged to obtain the splat normal. The splat size is an average distance from the depth sample to its neighbors. This fills most of the gaps that would otherwise appear due to the sparse set of depth samples in the DEP. The splats are texture mapped using the cube map faces (Figure 5-8).
5.3.2. Connected representation

An alternative to splatting is the connected representation of the DEP (Figure 5-9). It is built by triangulating in 2D the depth samples projected onto the faces of the cube map. A 3D triangle mesh is created by applying this connectivity data to the corresponding 3D depth samples (Figure 5-10). When rendering the model from a novel viewpoint, the 3D triangle mesh is texture-mapped with the cube map faces (Figure 5-11 and Figure 5-11).

During the acquisition, the 2D mesh is triangulated incrementally to accommodate the depth samples of the newly integrated frame. We use a Delaunay tree with logarithmic expected insertion time [DMT92, Devillers 92, BT93]. The implementation was obtained from [Delaunay].
The DEPs are compact and efficient data structures that enable modeling inside-looking-out scenes at interactive rates. Efficiency in registration and modeling comes from the fact that the scene is sampled from a single viewpoint. The DEPs
are comparable to color-only panoramas in terms of the acquisition cost, while they also capture the scene's geometry. The DEP provides a good raw model for a building section.
Figure 5-11 Depth samples are triangulated in 2D on the face of the panorama (orange) then the inferred connectivity is used to make the 3D mesh (blue).
CHAPTER 6. MODEL ASSEMBLY

Before a section model can be used in the building model, the raw DEP needs to be refined by proxy fitting and the section needs to be aligned with the building model constructed so far. The proxy models the major planes in the scene (walls, ceiling and floor) and is fitted to the point cloud, removing the need to acquire redundant dense depth samples. The extrapolated planes also allow the operator to register sections acquired from larger distance from each other, removing the need for overlapping point clouds for section registration.

6.1. Proxy fitting

We fit proxy geometry to the corridor or room DEPs. A corridor DEP has depth points for only a 2m band of the corridor tube. Outside of the band the DEP stores only color. We assume that the ceiling and floor are parallel, and that opposite walls are parallel to each other and perpendicular to the ceiling and floor. We fit four types of proxy geometry (see Figure 6-1): a rectangular box for rooms, an I section for a simple straight corridor piece, an L section for a corridor corner, and a T section for a corridor junction.

The fitting process starts by fitting floor and ceiling planes: the operator specifies the few points lying on the floor of the scene in the acquired model, and the plane is fitted through all the 3D samples inside the convex hull formed by the selected points. The same process is repeated to fit the ceiling plane. Then the operator specifies the lines in the color panorama where the walls intersect the ceiling plane. A downhill simplex search finds wall planes perpendicular to the ceiling plane closest to the lines specified by the operator. The walls are intersected with
floor and the ceiling to complete the proxy geometry. Orthographic texture is computed for each triangle from the cube map. The operator input is minimal, and it takes less than 1 minute to fit the section to the DEP.

In some cases the fitted section obstructs part of the scene: shelves behind the fitted wall of the room section, open door in the corridor section leading into a room, etc. The operator can specify a region of the section to be cut out. The remaining region is retriangulated automatically. The resulting triangles reuse the texture from the previous triangulation saving color reprojection costs.

Figure 6-1 Room section, and I, L, and T corridor sections.
6.2. Embedded detail

Building interiors are more complicated than the simple planar boxes of the proxies. Our system allows modeling the complex geometry inside rooms and enhancing the corridors with occasional geometric detail.

Recall that a DEP is a texture-mapped triangle mesh acquired from the acquisition point. The DEP is combined with a room box by first eliminating the DEP points that are close to the box faces. The threshold used in practice is 7cm.
Figure 6-3 Corridor section with embedded detail. Representing the entire wall with a single plane make the water fountain appear flat (top left). The operator selects points (top right) which are kept in the section model to convey correct motion parallax (bottom).

To make geometry recessed behind wall planes visible (e.g. water fountain in Figure 6-3), the operator cuts an opening in the wall in 2D, using a view from the center of the DEP. The points mapping to the opening are excluded from the planarity test. The second step is to sieve the connectivity data eliminating triangles for which all three vertices were discarded. This effectively flattens the part of the floor, ceiling, and walls visible in the DEP. The resulting section shows the overall room with flat walls, ceiling and floor, and with furniture and other geometric detail "sticking out" from these planes (Figure 6-2).
Many objects in the corridors can be truthfully modeled with texture data alone: doors, posters, ceiling lights. Occasionally, there are objects in the building corridors whose lack of geometry cannot be hidden with texture (e.g. benches, trash cans, fire extinguishers, and water fountains). They appear noticeably flat on the triangles of the corridor sections. Moreover, many of these are important for applications since they hinder access, or are useful in emergency response. During acquisition the operator samples the geometry of these objects by sweeping the laser beams repeatedly over them. After the section is fitted through the data, the operator selects region of the panorama with additional objects. We construct a plane cutting through the points in the selected region. The points then are triangulated in 2D by projecting onto the plane, color from the panorama is used to texture map the resulting 3D mesh connecting the points. The mesh is stored with the section, and when rendered the additional geometry provides correct visual clues in the novel views of the objects (Figure 6-3).

The triangulation of the sparse depth samples does not preserve correct depth discontinuities, resulting in visible artifacts in the rendered model: broken edges of the tables, shelves, monitors. The operator can improve the quality of the model, by introducing additional depth samples into the model. Two primitives can be fitted to regions specified by the operator: lines and planes. If the operator specifies a 2D segment, the corresponding 3D segment passing through the depth samples lying nearby in 2D is used to introduce new 3D samples computed at regular intervals (1 cm in practice). A similar process creates points on the boundary of the planar patch through points in the region specified by the operator. Line interpolation improves the straight edges of shelves, and plane interpolation is useful for tables, monitors and other planar surfaces. Typically, the operator spends 10-15 minutes improving the visual appearance of a section.
6.3. **Section registration**

Once the model of a section is complete, the newly completed section is registered with the previous ones, effectively attaching it to the model. We leverage same plane constraints characteristic to the indoor architectural environment. Shared floor and two non-parallel side walls are sufficient to automatically lock a room section in place. The case of corridor sections is more challenging. Consider a newly acquired I section that extends a corridor. The shared floor and parallel side walls elucidate only five of the six degrees of freedom. The translation along the corridor cannot be determined solely from the geometric data. We found that the automatic search for the remaining registration parameter (maximizing color match) often fails in practice due to very repetitive texture information inside the bare corridors (similar doors, ceiling lights at periodic intervals). We found that a robust and efficient approach is to rely on the operator to slide the new section into place, by finding the position which minimizes the color difference at the overlap region (Figure 6-4).

Once the section is registered, the floor, ceiling, and side walls it extends are re-triangulated to integrate the extension into the model. The texture is extended by merging the old and new texture. A smooth transition is obtained by blending at the overlap region. Using these four section types we can reconstruct complex floor plans, such as the floor shown in Figure 6-5. Note that the fitted sections allowed registering non overlapping point clouds.

After a floor has been constructed out of individual sections, it is added to the model of the building (Figure 6-6). Again, we make use of the geometric constraints. A floor is parallel to the floor below, shares a common wall (selected by the user) and is separated by a user specified distance between the two floors.
The total time to fit the proxy section to the point cloud, interpolate sparse depth samples inside rooms, add embed geometry to corridors, and register the section with the building model takes less than 10-15 minutes per DEP. While we have tried automatic methods for each proxy fitting and registration, they often failed due to sparse nature of the DEP and repetitive features inside the man-made environments. The operator input is minimal and allows performing these steps robustly.
Figure 6-4 Section registration: (top) original sections, (middle) sections after applying geometrical constraints, (bottom) connected sections.
Figure 6-5 Floor composed of 12 corridor and 4 room sections with the acquired point clouds.
Figure 6-6 Wireframe visualization of proxy geometry of building model.
A DEP samples the color and geometry of the scene from a single viewpoint, which greatly simplifies registration and modeling. A DEP provides a good visualization of the scene for viewpoints close to the acquisition viewpoint. However, the quality of the image degrades as the viewpoint moves away from the acquisition viewpoint and undersampled parts of the scene become visible (Figure 7-1). To increase the quality of the model one has to acquire data from several viewpoints.

Sampling the scene from several viewpoints while maintaining modeling efficiency is challenging. We have developed several preliminary methods for different types of scenes. We have promising results scanning structured scenes in handheld mode. Multi-viewpoint modeling of unstructured scenes is achieved by acquiring several DEPs of the same scene section and merging them in real time leveraging pixel level programmability of modern graphics hardware.
Figure 7-1 Depth enhanced panorama of plant shown from viewpoint near (top) and far (bottom) from the acquisition viewpoint.
7.1. Handheld structured scene acquisition

We acquire structured scenes using a hand held ModelCamera with the lasers turned on. The operator slowly moves the camera around the scene, sweeping the surfaces (Figure 7-2). Each frame containing the detected dots and dense color is processed in three steps:

1. fit the depth samples per frame with polynomial surfaces;
2. register consecutive frames using surfaces and color data;
3. merge the registered frames into a scene model.

Figure 7-2 Handheld acquisition of structured scenes (top), and evolving set of depth images shown with green dots (bottom).
7.1.1. Surface fitting

The dots in a frame are grouped into surfaces. Figure 7-3 shows the three surfaces fitted through the dots of the frame shown in Figure 4-7. The bottom four rows of dots lie on the couch backrest, the three right dots of the top three rows lie on the right wall, and the remaining dots lie on the left wall. Surface boundaries appear as depth discontinuities (couch/wall) or as curvature discontinuities (wall/wall). We construct a dot connectivity graph by scanning each row and column of dots for discontinuities. Adjacent dots are connected unless their second divided depth difference exceeds a threshold. We least squares fit a polynomial surface \( z = P(x, y) \) to the dots in each connected component. The dots are assigned surface normals according to the polynomials.

7.1.2. Frame registration

The registration task is to compute the motion of the ModelCamera from frame to frame. Given dense depth, registration is commonly performed with the iterative closest point algorithm [Besl 92, Rusinkiewicz 2002]. This algorithm does not work with sparse depth. Nor does it handle symmetric surfaces (explained below), which are the norm in structured scenes. We achieve fast, robust registration by first aligning the depths in the two frames then aligning the colors (Figure 7-4).
Depth registration
We perform depth registration by least-squares fitting laser dots in the new frame to old frame surfaces. The motion is linearized as $m(p) = t + p + r \times p$ with $t = (tx, ty, tz)$ the translation and $r = (rx, ry, rz)$ the angular velocity. An equation is generated for each dot that lies on the largest surface in both frames. The old
surface is linearized as $n(c-a) = 0$ with $n$ the surface normal, $c$ the new dot, and $a$ the old dot. The depth equation is $tn + r(c \times n) = n(a-c)$.

The depth equations form a system $Ax = b$ with $A$ a $k$-by-6 matrix, $x = (tx, ty, tz, rx, ry, rz)$ a 6 vector, and $b$ a $k$ vector. The system has a unique least-squares solution when $A$ has rank six. In structured scenes, the 49 dots provide ample equations, but symmetric surfaces generate linearly dependent equations. A surface is symmetric when it is invariant under translation along an axis or rotation about an axis. Examples are planes, surfaces of extrusion, surfaces of rotation, and spheres. The distance from the dots to the surface is constant when the camera performs these motions, so they cannot be computed from any amount of depth data.

We restrict the depth equations to a 3-dimensional subspace of $x$ that represents asymmetric motion. To identify this subspace, we classify the surface as cylindrical, spherical, or planar. We choose a coordinate system whose origin is the centroid of the surface, whose $z$ axis is the surface normal, and whose $x$ and $y$ axes are the major and minor principal axes (Figure 7-5). If the two principal curvatures are positive and are equal within a tolerance, the surface is classified as spherical and the motions assigned to depth registration are $tx, ty, tz$. If one curvature is positive and the other is zero, the surface is classified as cylindrical and the motions are $tx, tz,$ and $rx$. Otherwise the surface is classified as planar with motions $tz, rx,$ and $ry$. This case covers planes and asymmetric surfaces, which are well approximated by their tangent planes.
Color registration
We compute the symmetric xi's by minimizing a color error function. The error of a pixel in the new frame is the RGB distance between its color and the color where it projects in the old frame. The old color is computed by bilinear interpolation because the pixel projects at fractional coordinates. Small camera motions produce rapid, erratic changes in color error. We reduce the variability by convolving each frame with an 11-by-11 box filter. We then select a set of new pixels and minimize the sum of the squares of their errors by the downhill simplex method. This method is simple and does not require derivatives, which are expensive to compute.

The pixels are selected by scanning every kth row and column (we used k = 20) of the image and splitting them into segments. A segment is a maximal sequence of pixels that are dot free and that lie on a single surface. Dot pixels are excluded because their color comes from the lasers, rather than from the scene. The pixels are assigned depths by linear interpolation from the three nearest dots. They are projected into the old frame by incremental 3D warping [McMillan 1995, McMillan 1997]. Warped-image reconstruction is unnecessary for error evaluation, so this approach does not incur the full cost of IBR by 3D warping [Popescu 2003].
7.1.3. Modeling

The structured scene is modeled as a collection of images with per pixel depth (depth images). The depth images are created on demand as scanning progresses (Figure 7-6). We use depth images because they can be transformed and merged efficiently [Shade 1996, Popescu 2000, 2003]. Each registered frame is processed as follows. The region spanned by the dots is triangulated and transformed into a depth image by assigning a depth value to each color pixel in the region using the triangulation. The depth image of the new frame is merged into the depth images of the model. The better samples are retained. The quality metric is based on the sampling rate. Samples that project at the border between two depth images are repeated to provide overlap. The depth images are rendered efficiently as texture-mapped meshes to provide modeling feedback. The operator can select a visualization mode that highlights the parts of the model that were acquired below or above the desired sampling rate.

In the scanning experiments, we have observed the accumulation of the frame to frame registration error leading to the increasing drift in the subsequent frames. Due to the time constraints we postponed the investigation of the global registration algorithms that can alleviate this problem to the future research.

The modeling stage using depth images proved to be robust enough to register thousands of individual frames. For example, by positioning the camera on a tripod, and registering individual frames using panorama color registration method, we have been able to quickly merge thousands of frames into an armchair model with tens of depth images (Figure 7-7).
Figure 7-6 Couch armrest modeled with depth images created on demand.
Figure 7-7 Model obtained from a pre-registered sequence of 2300 frames

7.2. Run-time multiple DEP merging

If the desired view is close to the acquisition viewpoint, a single DEP produces high-quality images of the scene. If the desired view is considerably different
from the DEP acquisition view, the image quality degrades because of missing and undersampled surfaces. A wider range of views is supported by acquiring, registering and displaying several DEPs of the scene. The operator builds the first DEP as before, examines it for missing or poorly sampled surfaces, moves the ModelCamera to a second viewpoint, and starts building the second DEP. Once sufficient surfaces are acquired, the second DEP is registered with the first using three operator-specified point correspondences between the two DEPs. The system computes the rigid camera motion between the two DEPs, the acquisition of the second DEP resumes, and the two DEPs are visualized together.

### 7.2.1. Disconnected representation

The disconnected representation supports multiple DEP visualization without modification. Once a DEP is acquired, the ModelCamera is moved to a new location. During the acquisition of the second DEP, the already completed DEPs and the evolving DEP are rendered independently from each other in the splatting mode to guide the operator in completing the model (Figure 7-8).
7.2.2. Connected representation

To achieve higher quality visualization than a disconnected representation allows we developed a runtime per pixel sample selection algorithm to combine the best
parts of several DEPs. The straightforward z-buffered rendering of the multiple DEPs does not produce realistic novel views, since better sampled surfaces are often obscured by the worse sampled ones. To solve this problem, we wrote a GPU fragment program that selects the input sample for every pixel based on z and sampling rate (Figure 7-9 and Figure 7-12).

First, during a preprocessing step, we compute the sampling rate of each triangle as the inverse of the average length of its sides, normalized to 0..1 range. Triangles with sampling rates below a given threshold $SR$ are labeled as undersampled. $SR$ is established experimentally for each scene.

During runtime we select the three DEPs with acquisition viewpoints closest to the desired viewpoint. Each input DEP is rendered from the desired view into a separate high-resolution 2048x2048 pixel OpenGL color and z buffer. The sampling rate of the triangle is stored in the alpha channel. The separate color and z buffers are bound as texture maps and combined on per pixel basis using the fragment program. If a pixel is covered in at least one DEP by a triangle that is not undersampled, the closest such triangle is used (Figure 7-10). If an output pixel is covered by undersampled triangles in all of the DEPs, the algorithm selects the farthest undersampled surface (Figure 7-11).

The rendering frame rate depends on two main factors: the number of primitives in each input DEP and the target frame buffer resolution. The rendering of the static input DEPs is optimized by using the OpenGL compiled lists. The per pixel merging of the individual OpenGL color and depth buffers is sped up by using WGL_ARB_pbuffer and GL_ARB_multitexture extensions [SGI] which allow drawing into and combining multiple rendering contexts on the graphics card without transfers of the pixel buffers to/from main memory. We have achieved 5 fps rendering rate using the nVidia Quattro FX 3000 graphics card for a 512x512
output frame buffer and three input DEPs with more than 40K triangles each. The computation of the triangle sampling rate as preprocess took less than a second. We have observed visual artifacts in the undersampled regions, due to very coarse geometry of the DEPs in these areas. We believe that the ability to inspect the results immediately, as we have achieved, allows the operator to identify the undersampled regions, and improve the model appearance by scanning the scene from new locations.

The presented solution allowed us to render complex scenes, such as a plant in a corner, from large range of novel viewpoints without degradation in quality. At the same time this method has certain drawbacks: longer per-scene acquisition times, redundant data in the overlapping DEPs, low rendering frame rates. We believe that future research needs to address merging of depth enhanced panoramas acquired from multiple viewpoints to produce compact high quality models of unstructured scenes.
Figure 7-9 Input DEPs (top row), under sampling highlighted in red (bottom left), novel view rendered using sample selection based on sampling rate (bottom right).
Two input DEPs (red and blue) of a scene are combined to produce a novel view (green). The best sample for the desired view ray shown is the sample coming from the second DEP, although the sample coming from the first DEP is closer.

The desired view ray intersects undersampled triangles in all DEPs. We select the farthest sample, in this case the one coming from the second DEP. The farther sample invalidates the triangle coming from the first DEP, and indicates that the second DEP approximates the scene more closely.
Figure 7-12 Top row: two out of four input DEPs shown from their acquisition viewpoint. Middle row: each DEP shown from the desired view; the desired view is 151cm and 91cm away from the two acquisition points, respectively. Undersampling causes stretching. Bottom row: DEPs z-buffered together (left), and combined using GPU fragment shader which decides which sample to use based on both depth and sampling rate (right).
CHAPTER 8. RESULTS

To validate the system design and the modeling framework we have conducted an experiment. The goal was quickly modeling a real world office building including a hundred individual rooms in a week. Because of the logistics difficulties we could not get access to so many rooms, and the experiment was limited to the acquisition of corridors covering most of the 6 floors and 20 individual rooms.

We have placed the computer on a cart with wheels, together with the monitor to display the results in real time and an uninterruptible power supply to allow switching from one power outlet to another without shutting down the system. The acquisition device was placed on a sturdy tripod (Figure 8-1). Two people were needed to move this setup: one to push the cart, and another to carry the tripod with the camera. Only one person was needed during the actual scanning (Figure 1-3). The second person was identifying the next room to be scanned. The itinerary was finalized on the fly since we did not want to impose pre-established scanning times on those that had offices in the building. The 20 rooms were acquired over two days, with a single pass on each floor.

Corridors were captured by acquiring DEPs every 7 – 9 meters apart. The longest corridor section measures 36m and it was acquired with 5 DEPs from end to end. To minimize disruption to the normal activity in the building we did not cordon off the scene during scanning. The interruptions due to people moving through the corridors had a negligible impact due to the interactive nature of our acquisition pipeline. During the acquisition of a corridor DEP the cart was moved to remain outside of the field of view of the camera and therefore outside the panorama.
For room acquisition, the room had to be vacated for a total of 10 minutes. The device was positioned in the center of the room, and the cart was in the door frame. The door frame was cut out in the fitted box with operator input. On average, we spent 7/9 minutes acquiring depth and color for a corridor/room DEP. The longer acquisition times for the DEP are necessary to capture the more complex geometry.

The building model shown in Figure 1-5 contains 56 corridor sections and 20 individual rooms, spanning 6 floors. The corridor sections cover about 1,130 square meters of floor space. The room models cover 320 square meters of floor space. The original data for each room contains on average 110K depth samples. After discarding samples lying within 7 cm from the fitted section planes, 60K samples remain in each room, on average. For corridors sections we have acquired on average 38K samples, from which 5K samples were kept.

Section fitting takes on average 3 minutes, including computation of orthographic textures for the triangles of the section. It took about 2 minutes to register a pair of sections, and to recompute the shared textures, for a total per section time of less than 15 minutes. The proxies used in the model total less than 1,000 triangles. The fragmented geometry inside the room sections is modeled with 97K triangles per room, on average.

We have measured the dimensions of the longest corridor span on each floor. The average error in our model was 2.5%, although in one case the length of the corridor was off by 4.5%.

Our model is a set of texture mapped triangles saved in the VRML format. The model can be rendered with standard graphics APIs implemented in hardware (images throughout this chapter). The full resolution model contains ~2 million
triangles and over 2GB of textures. When the application desires to render the entire model a version with down sampled textures (4x4) and decimated geometry (90%) is used to enable interactive rates.

From our experience the hardware scanner proved to be robust and successfully acquired color and depth samples in varied environments, including points on specular objects, highly absorbent materials, or fragmented surfaces. In cases were depth samples could not be acquired reliably (computer monitors presented a problem), the correct depth could be interpolated using neighborhood samples and minimal user input. In many cases, the lack of depth samples was mitigated by the high quality texture maps.

During the acquisition and modeling we have noticed strong correlation between the geometric complexity of the scene and the scanning time. By relying on proxy fitting and not on dense depth sampling of walls we have been able to quickly scan most of the corridor sections, spending more time only in places with additional objects present. Likewise we have spent more time acquiring more depth samples inside cluttered offices, than in other, mostly empty rooms.
Figure 8-1 ModelCamera prototype 4 and the cart with PC can be easily moved to a new scanning location
Figure 8-2 Captured corridor section rendered from a point close to the acquisition location (above) and from an extreme view (below)
Figure 8-3 Rendering of the basement floor from novel locations
Figure 8-4 Captured room section rendered from two novel viewpoints
Figure 8-5 Model of an office from two viewpoints.
Figure 8-6 Room model rendered from two viewpoints with and without wireframe mesh
Figure 8-7 Basement floor model rendered from novel viewpoint
Figure 8-8 Corridor section with fragmented geometry.
We have described a system for fast large scale modeling of building interiors. The system relies on a custom acquisition device that reliably captures video frames enhanced with depth samples, and on interactive feedback to the operator who efficiently guides the acquisition.

We have refined the initial scanner design through several iterations, improving the configuration of the epipolar lines, improving the utilization of the video frame, and increasing the number of depth samples acquired in each frame. The latest design allows unambiguous detection of 11x11 laser dots by separating epipolar segments and acquires depth for objects that are challenging for prior active acquisition techniques, such as cluttered book shelves, plants with small dark green leaves, and clothes.

Our automated modeling pipeline with the steps of acquisition, registration, modeling, and visualization runs at 5 frames per second. The modular implementation of the pipeline allowed us to design and test different methods for registration and modeling depending on the type of the targeted scene.

We model unstructured scenes efficiently and effectively based on the Depth Enhanced Panorama, a novel modeling structure which is constructed incrementally at interactive rates. DEPs are built from frames with the same center of projection, which simplifies registration—frames are registered by stitching or using shaft encoders—and modeling—the depth samples are triangulated in 2D on the faces of a panorama centered at the acquisition.
viewpoint. The DEP handles frames with different exposure, handles uneven depth sampling resolution, and allows adding geometric detail at linear cost. The resulting texture-mapped triangle mesh enables computing realistic images of the scene at interactive rates using graphics hardware.

We take advantage of the simpler geometry of structured scenes with novel registration and modeling algorithms that allow scanning the scene in handheld mode at interactive rates. The registration algorithm relies solely on the depth and color data acquired to register the frames with different center of projection. The modeling algorithm merges the highly redundant frames into a set of disjoint depth images that are transformed in texture-mapped triangle meshes for realistic visualization at interactive rates. Like all pair-wise registration methods, our registration algorithm suffers from the accumulation of the frame to frame registration error.

The short acquisition time enables an interactive automated modeling pipeline which is substantially more efficient than pipelines based on lengthy acquisition of dense depth maps. Once the operator is effectively integrated in the modeling loop, modeling greatly benefits from the operator’s understanding of the scene. The operator monitors data acquisition and naturally aims the acquisition device towards the parts of the scene with complex geometry, maximizing scanning efficiency. A cluttered office is modeled in 15 minutes. Corridor segments are modeled even faster since a small number of depth samples is sufficient to estimate the planar primitives, which can be extended as long as the color sampling resolution is sufficient.

Sequences of sparse depth frames have great modeling power. The depth samples contributed by each frame quickly accumulate to capture a complex scene well. Additional information is inferred from the sparse point cloud and color of the DEP with operator’s help. Missing depth values along undersampled
edges of the furniture are inferred by fitting planes through operator-selected
regions. Similarly, walls, ceiling and floor in each scene are computed by plane
fitting and simplex search using operator input. The input needed to post-process
each DEP is small, utilizes operator’s knowledge of the scene. We have
experimented with existing and novel automatic methods that would remove the
need of operator input. However, the robustness of the manual approach makes
it preferable both in terms of efficiency and quality in the context of large scale
modeling.

A complex model is assembled from sections relying again on minimal operator
input for alignment. Geometric constraints provide all but one degree of freedom.
The operator easily slides the new section into place, monitoring the color
difference at the region of overlap. The resulting model consists of large triangles
for modeling the flat surfaces of the building interior such as ceiling, floor, and
walls, and is enhanced with high-resolution geometric detail where appropriate.
The high-quality color provides visual realism.

We validated our modeling system by acquiring a significant fraction of a large
building on campus. With only a minor interference with the normal activity in the
building, a team of two operators built what is, to the best of our knowledge and
that of our reviewers, the largest indoor scene reconstruction to date.

9.1. Future work
The current work could be extended in several ways. Low level development
could improve the usability of the various software tools, as well as making the
hardware more maneuverable. A tripod with wheels and battery power will allow
acquiring data more efficiently, by a single operator. Another improvement with
small cost and great benefit is modeling the materials frequently repeated
throughout a large building. In our case, the first candidate is the linoleum on the
corridor and room floors. The specularity is not negligible and accounting for it
will increase the realism of the walkthroughs by replacing the presently frozen highlights with correct, dynamic highlights. This also requires solving the problem of locating the light sources. Again, we plan to exploit the model regularity. Selecting two fluorescent lights should allow for the automatic instantiation of the remaining, evenly-spaced, lights.

With increased laser power output and longer baseline, the acquisition device could scan larger rooms, extending the system to other building types, such as museums and theaters.

The large number of identical objects inside man-made environments could be exploited by building and reusing a database of already scanned parts. This could greatly speed up acquisition and improve the resulting model by avoiding repeated modeling of chairs, tables, desk, and file cabinets.

Adding wireless connectivity will allow the second operator to fit the proxies and connect the sections remotely, from a model integration station. We do not foresee any major difficulty since the incremental updates to the color cube map and to the set of 3D points have a compact memory footprint. Rapid acquisition speed allowed by the ModelCamera would permit scanning with several acquisition devices in parallel, which could all be served by the same model integration station.

Finally, the ModelCamera could be extended to acquire video frames from multiple centers of projection by replacing the pan and tilt bracket with a mechanical tracking device like an arm presently used for digitization in reverse engineering. Freely translating the acquisition device would allow the operator to easily capture occluded regions of the scene, improving the quality of final model. The tracker will provide registration so the research effort will only have to address modeling of unstructured scenes from frames with different centers of
projection. This will set the stage for the ultimate goal of the ModelCamera project, which is to make 3D modeling as simple, intuitive, and inexpensive as home video.
LIST OF REFERENCES
3DV Systems. http://www.3dvsystems.com


SGI OpenGL extensions specification.
http://oss.sgi.com/projects/oglsample/registry/


VITA

Gleb Bahmutov was born on January 25, 1980 in Chișinău, Moldova. After receiving his higher secondary education, he attended Purdue University at Indianapolis and received his Bachelor of Science in 2000. During the last year of his undergraduate studies he has worked both part- and full-time as a software engineer in a start-up company ObjectBuilders Inc. In August, 2000, he started graduate studies at Computer Science Department, Purdue University at West Lafayette. After attending a class led by new faculty member Dr. Voicu Popescu, he became interested in computer graphics field. After finishing the Ph.D. program, Gleb Bahmutov joined an industrial company as a researcher in 3D acquisition and modeling area.


Gleb Bahmutov was born on January 25, 1980 in Chișinău, Moldova. After receiving his higher secondary education, he attended Purdue University at Indianapolis and received his Bachelor of Science in 2000. During the last year of his undergraduate studies he has worked both part- and full-time as a software engineer in a start-up company ObjectBuilders Inc. In August, 2000, he started graduate studies at Computer Science Department, Purdue University at West Lafayette. After attending a class led by new faculty member Dr. Voicu Popescu, he became interested in computer graphics field. After finishing the Ph.D. program, Gleb Bahmutov joined an industrial company as a researcher in 3D acquisition and modeling area.