Modeling and Rendering Architecture from Photographs

by

Paul Ernest Debevec

B.S.E. (University of Michigan at Ann Arbor) 1992
B.S. (University of Michigan at Ann Arbor) 1992

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science

in the

GRADUATE DIVISION
of the
UNIVERSITY of CALIFORNIA at BERKELEY

Committee in charge:

Professor Jitendra Malik, Chair
Professor John Canny
Professor David Wessel

Fall 1996

The dissertation of Paul E. Debevec is approved:

Jitendra Malik December 19, 1996

Chair Date

John Canny Dec 19, 1996

Date

David Wessel December 18, 1996

Date

University of California at Berkeley

1996

Modeling and Rendering Architecture from Photographs

by

Paul Ernest Debevec

Abstract

Modeling and Rendering Architecture from Photographs

by

Paul Ernest Debevec

Doctor of Philosophy in Computer Science

University of California at Berkeley

Professor Jitendra Malik, Chair

Imagine visiting your favorite place, taking a few pictures, and then turning those pictures into a photorealisic three-dimensional computer model. The work presented in this thesis combines techniques from computer vision and computer graphics to make this possible. The applications range from architectural planning and archaeological reconstructions to virtual environments and cinematic special effects.

This thesis presents an approach for modeling and rendering existing architectural scenes from sparse sets of still photographs. The modeling approach, which combines both geometry-based and image-based techniques, has two components. The first component is an interactive *photogrammetric modeling* method which facilitates the recovery of the basic geometry of the photographed scene. The photogrammetric modeling approach is effective, convenient, and robust because it exploits the constraints that are characteristic of architectural scenes. The second component is a *model-based* stereo algorithm, which recovers how the real scene deviates from the basic model. By mak-

ing use of the model, this new technique robustly recovers accurate depth from widely-spaced image pairs. Consequently, this approach can model large architectural environments with far fewer photographs than current image-based modeling approaches. For producing renderings, this thesis presents *view-dependent texture mapping*, a method of compositing multiple views of a scene that better simulates geometric detail on basic models.

This approach can be used to recover models for use in either geometry-based or image-based rendering systems. This work presents results that demonstrate the approach's ability to create realistic renderings of architectural scenes from viewpoints far from the original photographs. This thesis concludes with a presentation of how these modeling and rendering techniques were used to create the interactive art installation *Rouen Revisited*, presented at the SIGGRAPH '96 art show.

Professor Jitendra Malik
Dissertation Committee Chair

To Herschel

1983 - 1996

Contents

List of Figures

List of Tables

Acknowledgements

There are many people without whom this work would not have been possible. I would first like to thank my advisor, Jitendra Malik, for taking me on as his student and allowing me to pursue my research interests, and for his support and friendship during the course of this work. He is, without question, the most enjoyable person I could imagine having as a research advisor. I would also like to thank my collaborator C. J. Taylor, whose efforts and expertise in structure from motion formed the basis of our photogrammetric modeling system.

I would like to give a special thanks to Golan Levin, with whom I worked to create the *Rouen Revisited* art installation, for a very enjoyable and successful collaboration. His creativity and diligence helped inspire me to give my best efforts to the project. And I would like to thank David Liddle and Paul Allen of Interval Research Corporation for allowing *Rouen Revisited* to happen. It is, quite literally, a beautiful showcase for the methods developed in this thesis.

Also at Interval, I would like to thank Michael Naimark and John Woodfill for their inspiring successes with image-based modeling and rendering, and for demonstrating by example that you don't have to choose between having a career in art and a career in technology.

There are several other people whom I would like to thank for their support and encouragement. My housemates Judy Liu and Jennifer Brunson provided much-needed support during a variety of deadline crunches. I would like to thank Professors Carlo Séquin and David Forsyth for their interest in this work and their valuable suggestions. And I must especially thank Tim Hawkins for his unwavering support as well as his frequent late-night help discussing and revising the work that has gone into this thesis. Not only has Tim helped shape the course of this work, but due to his assistance more than *ninety-eight percent* of the sentences in this thesis contain verbs.

I would also like to thank my committee members John Canny and David Wessel for helping make this thesis happen despite somewhat inconvenient circumstances. Professor Wessel is, at the time of this writing, looking over a draft of this work in a remote farm house in France, just an hour south of Rouen.

I would also like to thank the sponsors of this research: the National Science Foundation Graduate Research Fellowship program, Interval Research Corporation, the California MICRO program, and JSEP contract F49620-93-C-0014.

And of course my parents.

Chapter 1

Introduction

Architecture is the art that we walk amongst and live within. It defines our cities, anchors our memories, and draws us forth to distant lands. Today, as they have for millenia, people travel throughout the world to marvel at architectural environments from Teotihuacan to the Taj Mahal. As the technology for experiencing immersive virtual environments develops, there will be a growing need for interesting virtual environments to experience. Our intimate relationship with the buildings around us attests that architectural environments — especially ones of cultural and historic importance — will provide some of the future's most compelling virtual destinations. As such, there is a clear call for a method of conveniently building photorealistic models of existing and historic architecture.

Already, efforts to build computer models of architectural scenes have produced many interesting applications in computer graphics; a few such projects are shown in Fig. 1.1. Unfortunately, the traditional methods of constructing models (Fig. 1.2a) of existing architecture, in which a modeling program is used to manually position the elements of the scene, have several drawbacks. First,

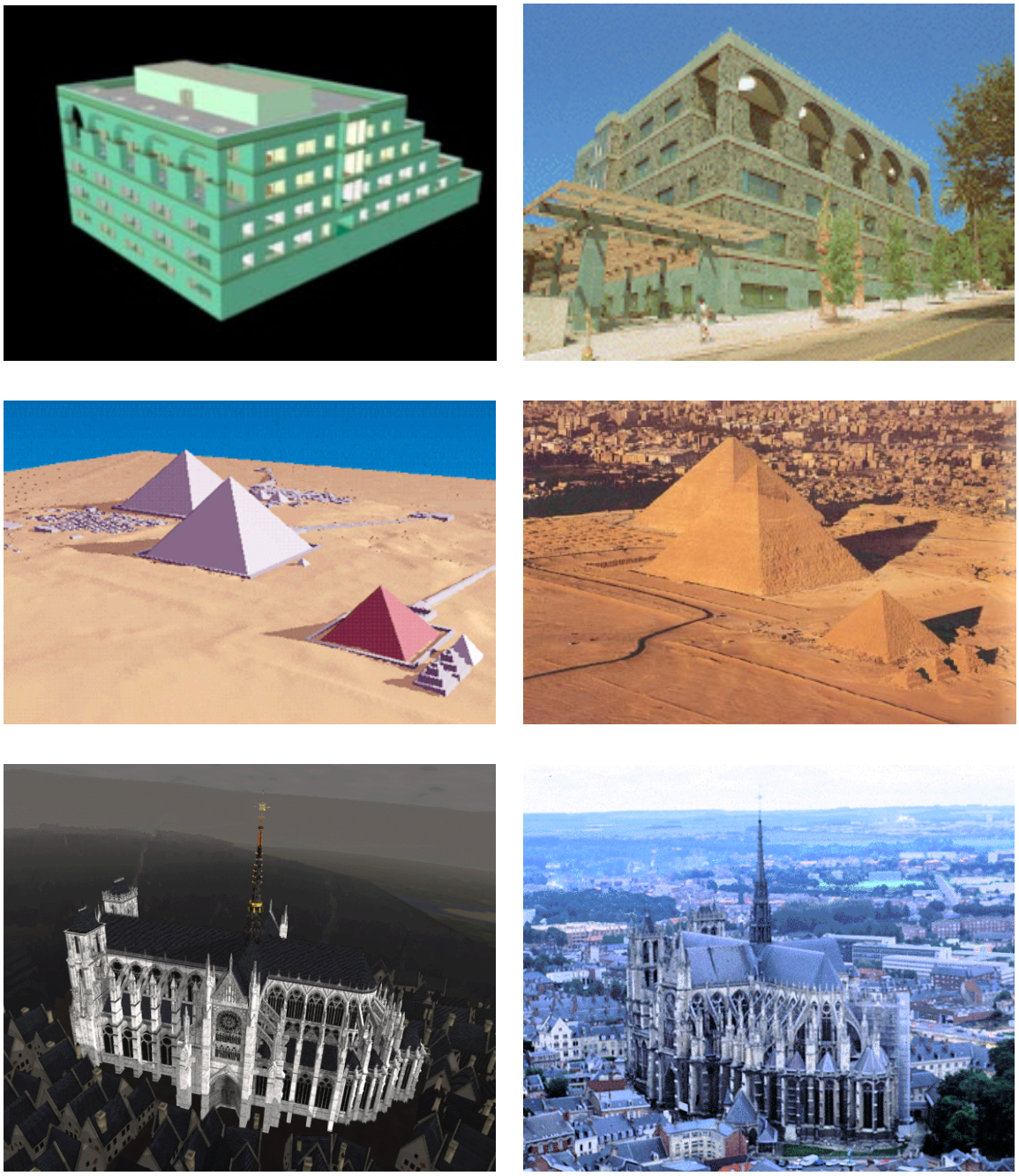


Figure 1.1: Three ambitious projects to model architecture with computers, each presented with a rendering of the computer model and a photograph of the actual architecture. Top: Soda Hall Walk-thru Project [49, 15], University of California at Berkeley. Middle: Giza Plateau Modeling Project, University of Chicago. Bottom: Virtual Amiens Cathedral, Columbia University. Using traditional modeling techniques (Fig. 1.2a), each of these models required many person-months of effort to build, and although each project yielded enjoyable and useful renderings, the results are qualitatively very different from actual photographs of the architecture.

the process is extremely labor-intensive, typically involving surveying the site, locating and digitizing architectural plans (if available), or converting existing CAD data (again, if available). Second, it is difficult to verify whether the resulting model is accurate. Most disappointing, though, is that the renderings of the resulting models are noticeably computer-generated; even those that employ liberal texture-mapping generally fail to resemble real photographs. As a result, it is very easy to distinguish the computer renderings from the real photographs in Fig. 1.1.

Recently, creating models directly from photographs has received increased interest in both computer vision and in computer graphics under the title of image-based modeling and rendering. Since real images are used as input, such an image-based system (Fig. 1.2c) has an advantage in producing photorealistic renderings as output. Some of the most promising of these systems ([24, 31, 27, 44, 37], see also Figs. 1.3 and 1.4) employ the computer vision technique of computational stereopsis to automatically determine the structure of the scene from the multiple photographs available. As a consequence, however, these systems are only as strong as the underlying stereo algorithms. This has caused problems because state-of-the-art stereo algorithms have a number of significant weaknesses; in particular, the photographs need to have similar viewpoints for reliable results to be obtained. Because of this, current image-based techniques must use many closely spaced images, and in some cases employ significant amounts of user input for each image pair to supervise the stereo algorithm. In this framework, capturing the data for a realistically renderable model would require an impractical number of closely spaced photographs, and deriving the depth from the photographs could require an impractical amount of user input. These concessions to the weakness of stereo algorithms would seem to bode poorly for creating large-scale, freely navigable virtual environments from photographs.

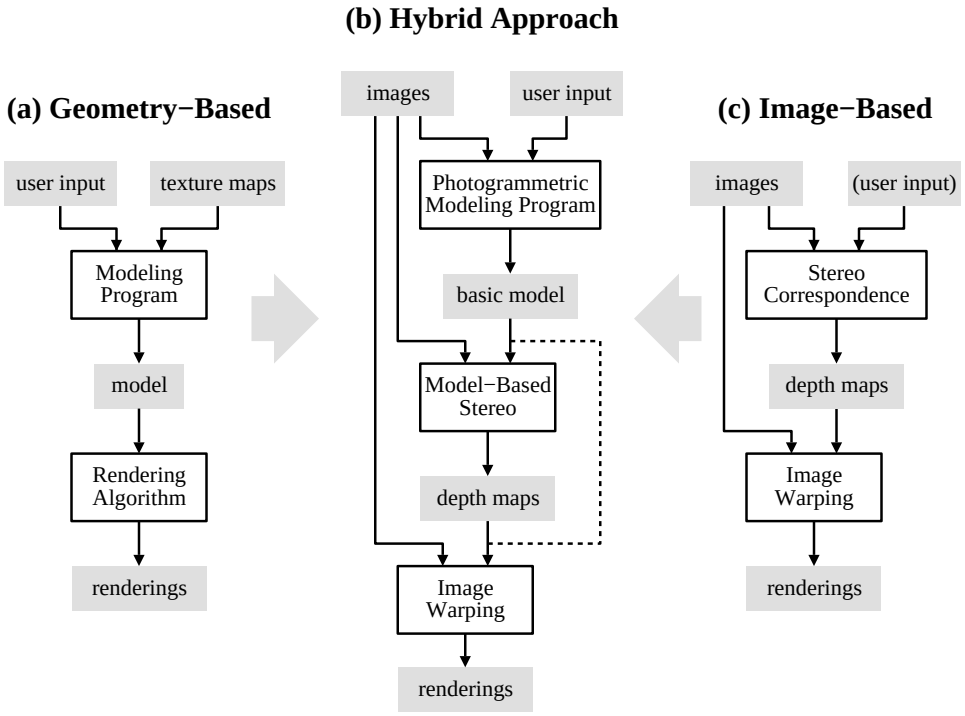


Figure 1.2: Schematic of how our hybrid approach combines geometry-based and image-based approaches to modeling and rendering architecture from photographs. The geometry-based approach illustrated places the majority of the modeling task on the user, whereas the image-based approach places the majority of the task on the computer. Our method divides the modeling task into two stages, one that is interactive, and one that is automated. The dividing point we have chosen capitalizes on the strengths of both the user and the computer to produce the best possible models and renderings using the fewest number of photographs. The dashed line in the geometry-based schematic indicates that images may optionally be used in a modeling program as texture-maps. The dashed line in the image-based schematic indicates that in some systems user input is used to initialize the stereo correspondence algorithm. The dashed line in the hybrid schematic indicates that view-dependent texture-mapping (as discussed in Chapter 6) can be used without performing stereo correspondence.

Figure 1.3: The Immersion ’94 stereo image sequence capture rig, being operated by Michael Naimark of Interval Research Corporation. Immersion ’94 was one project that attempted to create navigable, photorealistic virtual environments from photographic data. The stroller supports two identical 16mm movie cameras, and has an encoder on one wheel to measure the forward motion of the rig. The cameras are motor-driven and can be programmed to take pictures in synchrony at any distance interval as the camera rolls forward. For much of the work done for the *See Banff!* project, the forward motion distance between acquired stereo pairs was one meter. Photo by Louis Psihoyos-Matrix reprinted from the July 11, 1994 issue of Fortune Magazine.

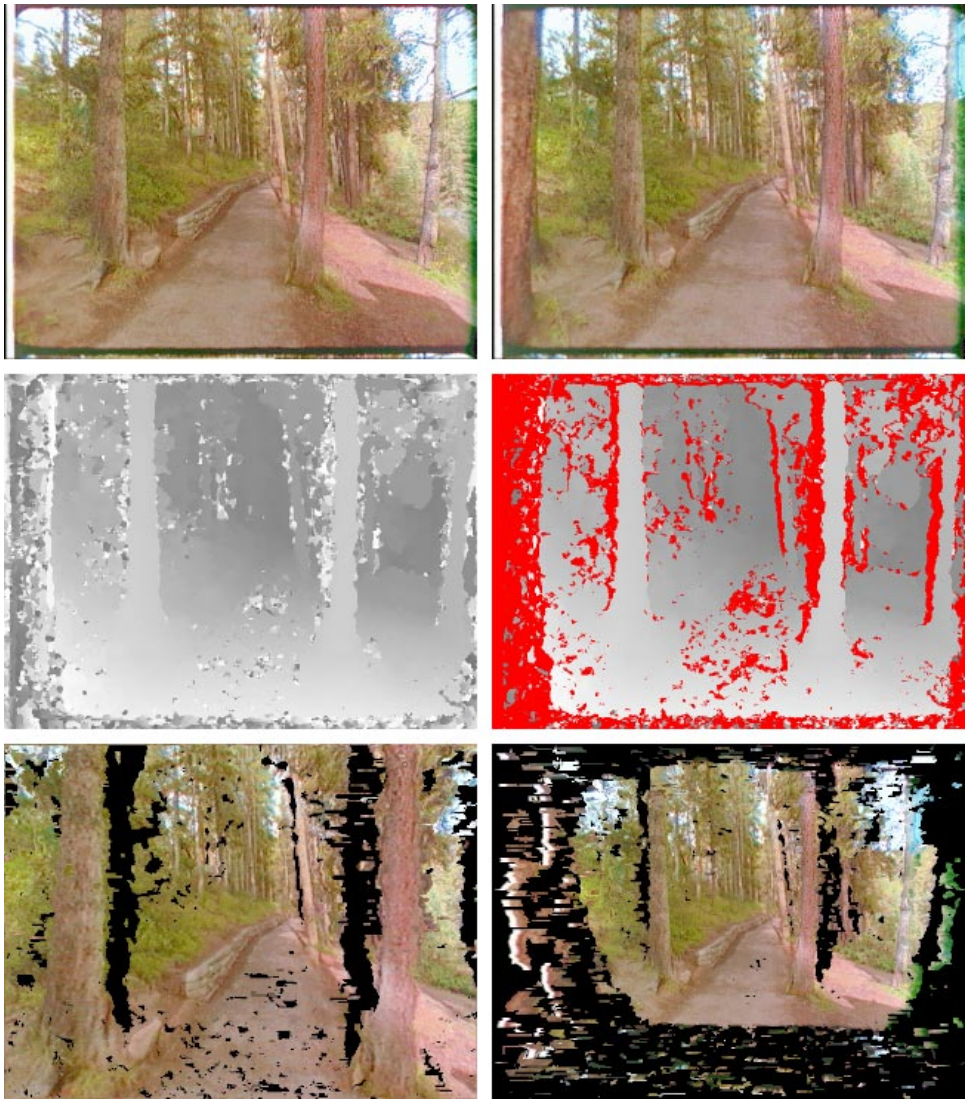


Figure 1.4: The Immersion ’94 image-based modeling and rendering (see Fig. 1.2c) project. The top two photos are a stereo pair (reversed for cross-eyed stereo viewing) taken with in Canada’s Banff National. The film frame was overscanned to assist in image registration. The middle left photo is a stereo disparity map produced by a parallel implementation of the Zabih-Woodfill stereo algorithm [59]. To its right the map has been processed using a left-right consistency check to invalidate regions where running stereo based on the left image and stereo based on the right image did not produce consistent results. Below are two virtual views generated by casting each pixel out into space based on its computed depth estimate, and reimaging the pixels into novel camera positions. On the left is the result of virtually moving one meter forward, on the right is the result of virtually moving one meter backward. Note the dark de-occluded areas produced by these virtual camera moves; these areas were not seen in the original stereo pair. In the Immersion ’94 animations, these regions were automatically filled in from neighboring stereo pairs.

The research presented here aims to make the process of modeling architectural scenes more convenient, more accurate, and more photorealistic than the methods currently available. To do this, we have developed a new approach that draws on the strengths of both traditional geometry-based and novel image-based methods, as illustrated in Fig. 1.2b. The result is that our approach to modeling and rendering architecture requires only a sparse set of photographs and can produce realistic renderings from arbitrary viewpoints. In our approach, a basic geometric model of the architecture is recovered semi-automatically with an easy-to-use photogrammetric modeling system (Chapter 5), novel views are created using view-dependent texture mapping (Chapter 6), and additional geometric detail can be recovered automatically through model-based stereo correspondence (Chapter 7). The final images can be rendered with current image-based rendering techniques or with traditional texture-mapping hardware. Because only photographs are required, our approach to modeling architecture is neither invasive nor does it require architectural plans, CAD models, or specialized instrumentation such as surveying equipment, GPS sensors or range scanners.

Chapter 2

Background and Related Work

The process of recovering 3D structure from 2D images has been a central endeavor within computer vision, and the process of rendering such recovered structures is an emerging topic in computer graphics. Although no general technique exists to derive models from images, several areas of research have provided results that are applicable to the problem of modeling and rendering architectural scenes. The particularly relevant areas reviewed here are: Camera Calibration, Structure from Motion, Shape from Silhouette Contours, Stereo Correspondence, and Image-Based Rendering.

2.1 Camera calibration

Recovering 3D structure from images becomes a simpler problem when the images are taken with *calibrated* cameras. For our purposes, a camera is said to be *calibrated* if the mapping between image coordinates and directions relative to the camera center are known. However, the position of the camera in space (i.e. its translation and rotation with respect to world coordinates) is not necessarily known. An excellent presentation of the algebraic and matrix representations of

perspective cameras may be found in [13].

Considerable work has been done in both photogrammetry and computer vision to calibrate cameras and lenses for both their perspective intrinsic parameters and their distortion patterns. Some successful methods include [52], [12], and [11]. While there has been recent progress in the use of uncalibrated views for 3D reconstruction [14], this method does not consider non-perspective camera distortion which prevents high-precision results for images taken through real lenses. In our work, we have found camera calibration to be a straightforward process that considerably simplifies the problem of 3D reconstruction. Chapter 4 provides a more detailed overview of the issues involved in camera calibration and presents the camera calibration process used in this work.

2.2 Structure from motion

Given the 2D projection of a point in the world, its position in 3D space could be anywhere on a ray extending out in a particular direction from the camera's optical center. However, when the projections of a sufficient number of points in the world are observed in multiple images from different positions, it is mathematically possible to deduce the 3D locations of the points as well as the positions of the original cameras, up to an unknown factor of scale.

This problem has been studied in the area of photogrammetry for the principal purpose of producing topographic maps. In 1913, Kruppa [23] proved the fundamental result that given two views of five distinct points, one could recover the rotation and translation between the two camera positions as well as the 3D locations of the points (up to a scale factor). Since then, the problem's mathematical and algorithmic aspects have been explored starting from the fundamental work of Ullman [54] and Longuet-Higgins [25], in the early 1980s. Faugeras's book [13] overviews the state

of the art as of 1992. So far, a key realization has been that the recovery of structure is very sensitive to noise in image measurements when the translation between the available camera positions is small.

Attention has turned to using more than two views with image stream methods such as [50] or recursive approaches [2]. Tomasi and Kanade [50] (see Fig. 2.1) showed excellent results for the case of orthographic cameras, but direct solutions for the perspective case remain elusive. In general, linear algorithms for the problem fail to make use of all available information while nonlinear optimization methods are prone to difficulties arising from local minima in the parameter space. An alternative formulation of the problem by Taylor and Kriegman [47] (see Fig. 2.2) uses lines rather than points as image measurements, but the previously stated concerns were shown to remain largely valid. For purposes of computer graphics, there is yet another problem: the models recovered by these algorithms consist of sparse point fields or individual line segments, which are not directly renderable as solid 3D models.

In our approach, we exploit the fact that we are trying to recover geometric models of architectural scenes, not arbitrary three-dimensional point sets. This enables us to include additional constraints not typically available to structure from motion algorithms and to overcome the problems of numerical instability that plague such approaches. Our approach is demonstrated in a useful interactive system for building architectural models from photographs (Chapter 5.)

2.3 Shape from silhouette contours

Some work has been done in both computer vision and computer graphics to recover the shape of objects from their silhouette contours in multiple images. If the camera geometry is known for each image, then each contour defines an infinite, cone-shaped region of space within which the

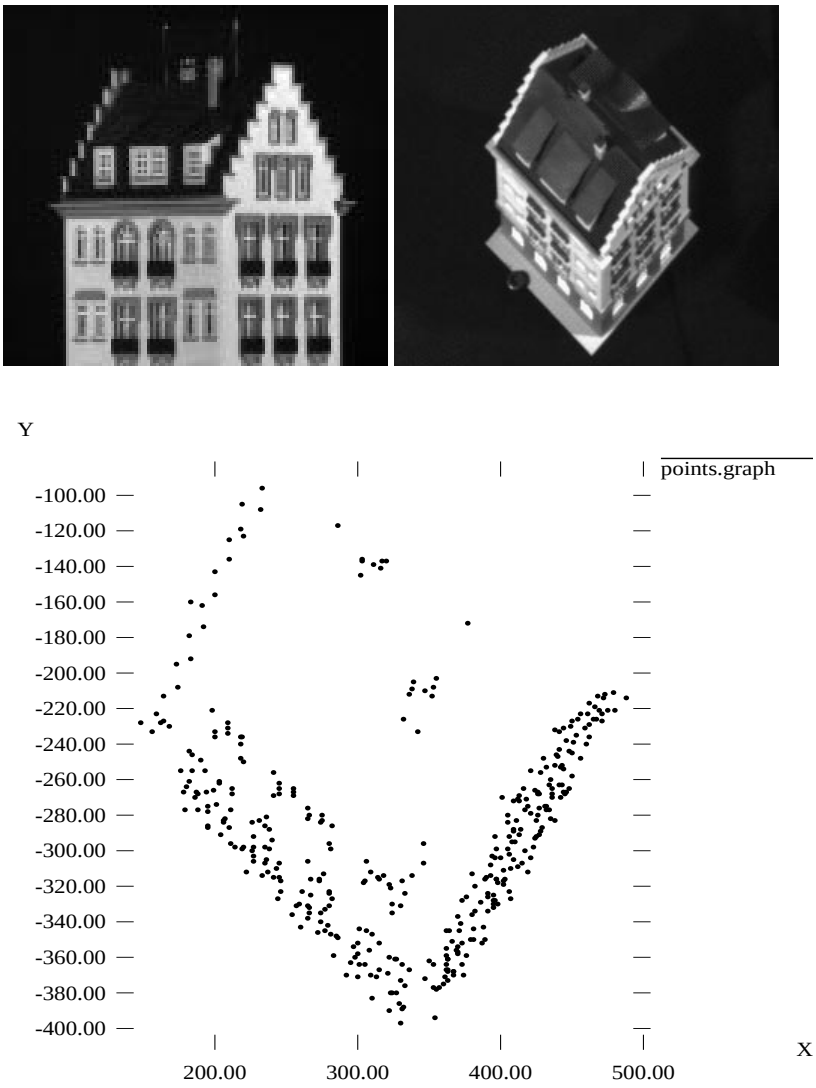


Figure 2.1: Images from the 1992 Tomasi-Kanade structure from motion paper [50]. In this paper, feature points were automatically tracked in an image sequence of a model house rotating. By assuming the camera was orthographic (which was approximated by using a telephoto lens), they were able to solve for the 3D structure of the points using a linear factorization method. The above left picture shows a picture from the original sequence, the above right picture shows a second image of the model from above (not in the original sequence), and the plot below shows the 3D recovered points from the same camera angle as the above right picture. Although an elegant and fundamental result, this approach is not directly applicable to real-world scenes because real camera lenses (especially those typically used for architecture) are too wide-angle to be approximated as orthographic.

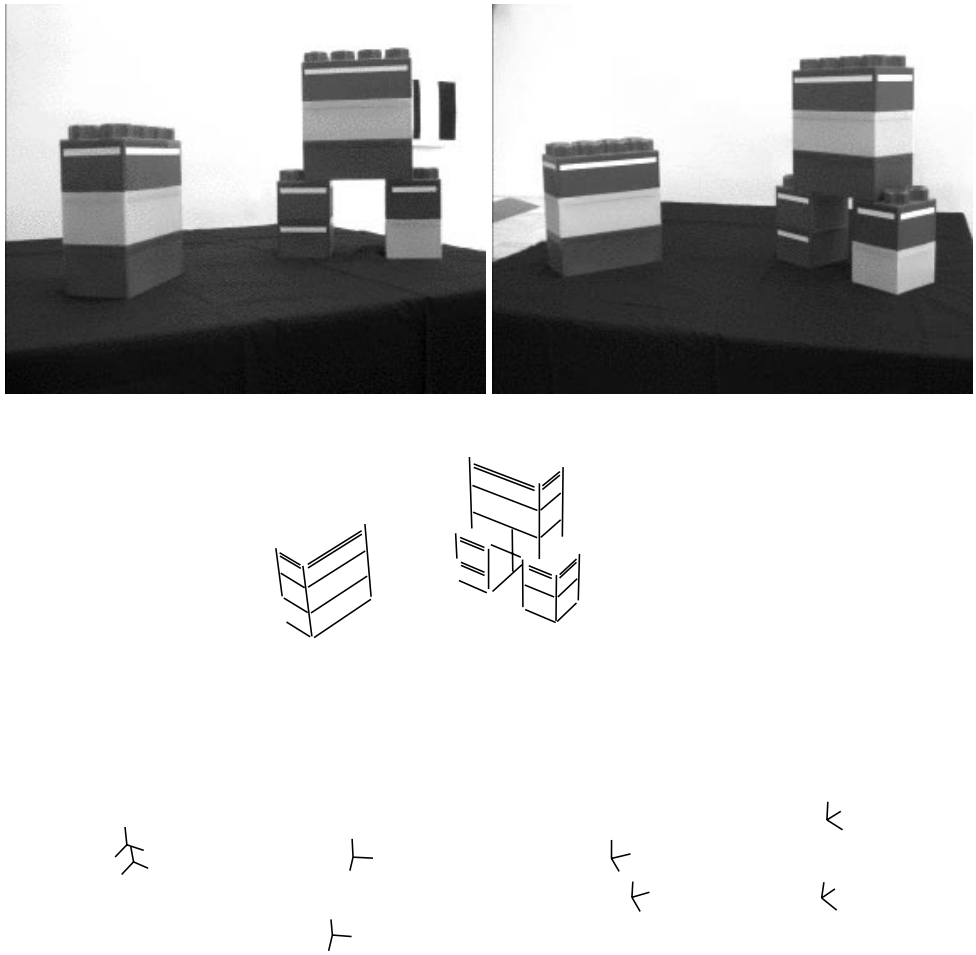


Figure 2.2: Images from the 1995 Taylor-Kriegman structure from motion paper [47]. In this work, structure from motion is recast in terms of line segments rather than points. A principal benefit of this is that line features are often more easily located in architectural scenes than point features. Above are two of eight images of a block scene; edge correspondences among the images were provided to the algorithm by the user. The algorithm then employed a nonlinear optimization technique to solve for the 3D positions of the line segments as well as the original camera positions, show below. This work used calibrated cameras, but allowed a full perspective model to be used in contrast to Tomasi and Kanade [50]. However, the optimization technique was prone to getting caught in local minima unless good initial estimates of the camera orientations were provided. This work was extended to become the basis of the photogrammetric modeling method presented in Chapter 5.

object must lie. An estimate for the geometry of the object can thus be obtained by intersecting multiple such regions from different images. As a greater variety of views of the object are used, this technique can eventually recover the ray hull[1] of the object. A simple version of the basic technique was demonstrated in [8], shown in Fig. 2.3. In this project, three nearly orthographic photographs of a car were used to carve out its shape, and the images were mapped onto this geometry to produce renderings. Although just three views were used, the recovered shape is close to the actual shape because the views were chosen to align with the boxy geometry of the object. A project in which a continuous stream of views was used to reconstruct object geometry is presented in [45, 44]; see also Fig. 2.4. A similar silhouette-based technique was used to provide an approximate estimate of object geometry to improve renderings in the Lumigraph image-based modeling and rendering system [16].

In modeling from silhouettes, qualitatively better results can be obtained for curved objects by assuming that the object surface normal is perpendicular to the viewing direction at every point of the contour. Using this constraint, [43] developed a surface fitting technique to recover curved models from images.

In general, silhouette contours can be used effectively to recover approximate geometry of individual objects, and the process can be automated if there is known camera geometry and the objects can be automatically segmented out of the images. Silhouette contours can also be used very effectively to recover the precise geometry of surfaces of revolution in images. However, for the general shape of an arbitrary building that has many sharp corners and concavities, silhouette contours alone can not provide adequately accurate model geometry.

[1]The ray hull of an object is the complement of the union of all rays in space which do not intersect the object. The ray hull can capture some forms of object concavities, but not, in general, complicated concave structure.

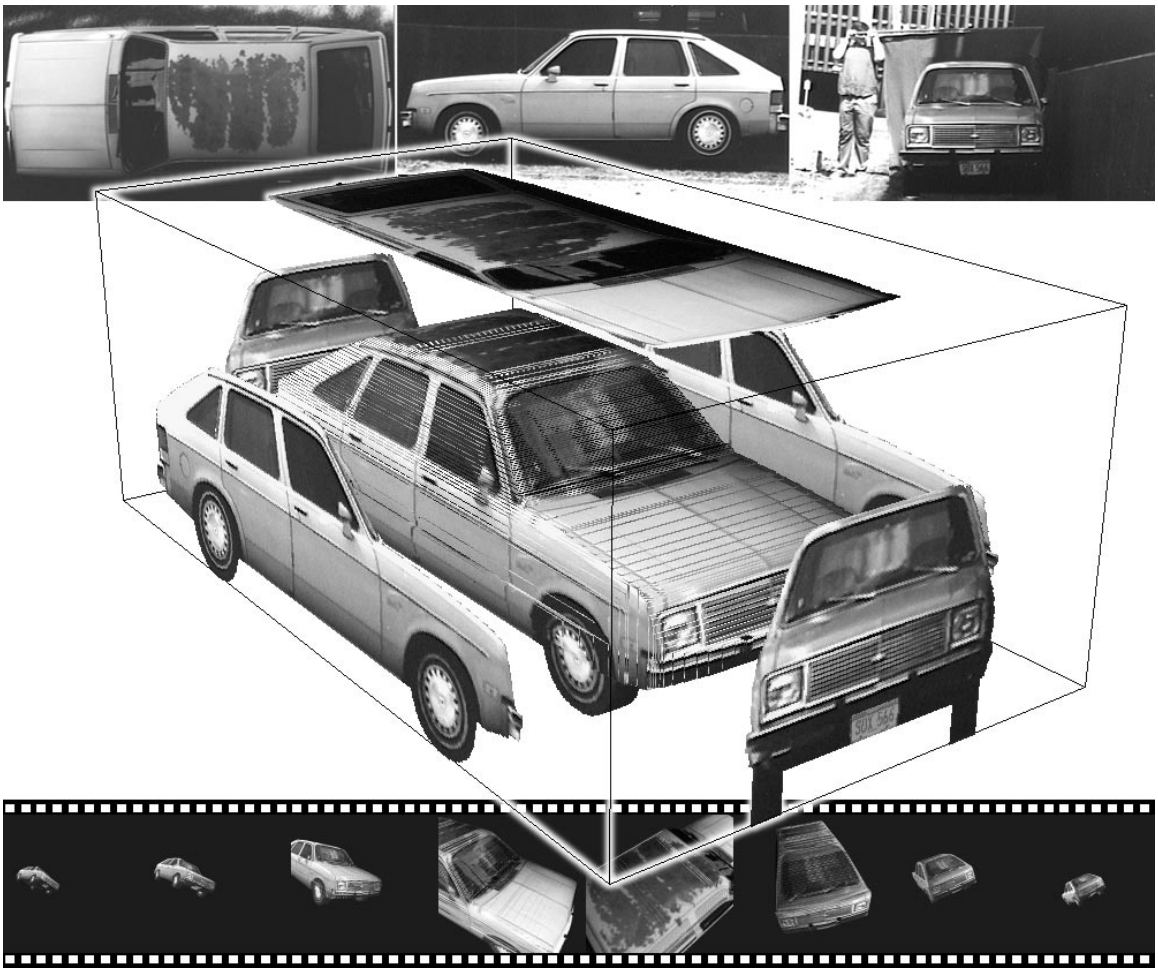


Figure 2.3: Images from the 1991 Chevette Modeling project [8]. The top three images show pictures of the 1980 Chevette photographed with a 210mm lens from the top, side, and front. The Chevette was semi-automatically segmented from each image, and these images were then registered with each other approximating the projection as orthographic. The registered photographs are shown placed in proper relation to each other on the faces of a rectangular box in the center of the figure. The shape of the car is then carved out from the box volume by perpendicularly sweeping each of the three silhouettes like a cookie-cutter through the box volume. The recovered volume (shown inside the box) is then textured-mapped by projecting the original photographs onto it. The bottom of the figure shows a sampling of frames from a synthetic animation of the car flying across the screen. Although (and perhaps because) the final model has flaws resulting from specularities, missing concavities, and imperfect image registration, it unequivocally evokes an uncanny sense of the actual vehicle.

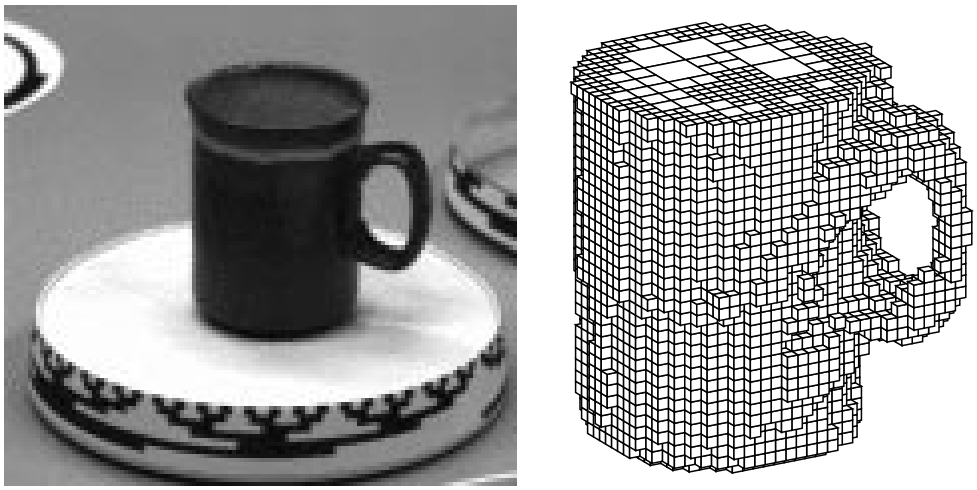


Figure 2.4: Images from a silhouette modeling project by Rick Szeliski [45, 44]. The cup was videotaped on a rotating platform (left), and the extracted contours from this image sequence were used to automatically recover the shape of the cup (right).

Although not adequate for general building shapes, silhouette contours could be useful in recovering the approximate shapes of trees, bushes, and topiary in architectural scenes. Techniques such as those presented in [36] could then be used to synthesize detailed plant geometry to conform to the shape and type of the original flora. This technique would seem to hold considerably more promise for practically recovering plant structure than trying to reconstruct the position and coloration of each individual leaf and branch of every tree in the scene.

2.4 Stereo correspondence

The geometrical theory of structure from motion assumes that one is able to solve the *correspondence* problem, which is to identify the points in two or more images that are projections of the same point in the world. In humans, corresponding points in the two slightly differing images on the retinas are determined by the visual cortex in the process called binocular stereopsis. Two terms used in reference to stereo are *baseline* and *disparity*. The baseline of a stereo pair is the distance

between the camera locations of the two images. Disparity refers to the difference in image location between corresponding features in the two images, which is projectively related to the depth of the feature in the scene.

Years of research (e.g. [3, 10, 17, 22, 26, 32, 35]) have shown that determining stereo correspondences by computer is difficult problem. In general, current methods are successful only when the images are similar in appearance, as in the case of human vision, which is usually obtained by using cameras that are closely spaced relative to the objects in the scene. As the distance between the cameras (often called the *baseline*) increases, surfaces in the images exhibit different degrees of foreshortening, different patterns of occlusion, and large disparities in their locations in the two images, all of which makes it much more difficult for the computer to determine correct stereo correspondences. To be more specific, the major sources of difficulty include:

1. **Foreshortening**. Surfaces in the scene viewed from different positions will be foreshortened differently in the images, causing the image neighborhoods of corresponding pixels to appear dissimilar. Such dissimilarity can confound stereo algorithms that use local similarity metrics to determine correspondences.

2. **Occlusions**. Depth discontinuities in the world can create half-occluded regions in an image pair, which also poses problems for local similarity metrics.

3. **Lack of Texture**. Where there is an absence of image intensity features it is difficult for a stereo algorithm to correctly find the correct match for a particular point, since many point neighborhoods will be similar in appearance.

Unfortunately, the alternative of improving stereo correspondence by using images taken from nearby locations has the disadvantage that computing depth becomes very sensitive to noise in

image measurements. Since depth is computed by taking the inverse of disparity, image pairs with small disparities tend to give rise to noisy depth estimates. Geometrically, depth is computed by triangulating the position of a matched point from its imaged position in the two cameras. When the cameras are placed close together, this triangle becomes very narrow, and the distance to its apex becomes very sensitive to the angles at its base. Noisy depth estimates mean that novel views will become visually unconvincing very quickly as the virtual camera moves away from the original viewpoint.

Thus, computing scene structure from stereo leaves us with a conundrum: image pairs with narrow baselines (relative to the distance of objects in the scene) are similar in appearance and make it possible to automatically compute stereo correspondences, but give noisy depth estimates. Image pairs with wide baselines can give very accurate depth localization for matched points, but the images usually exhibit large disparities, significant regions of occlusion, and different forms of foreshortening which makes it very difficult to automatically determine correspondences.

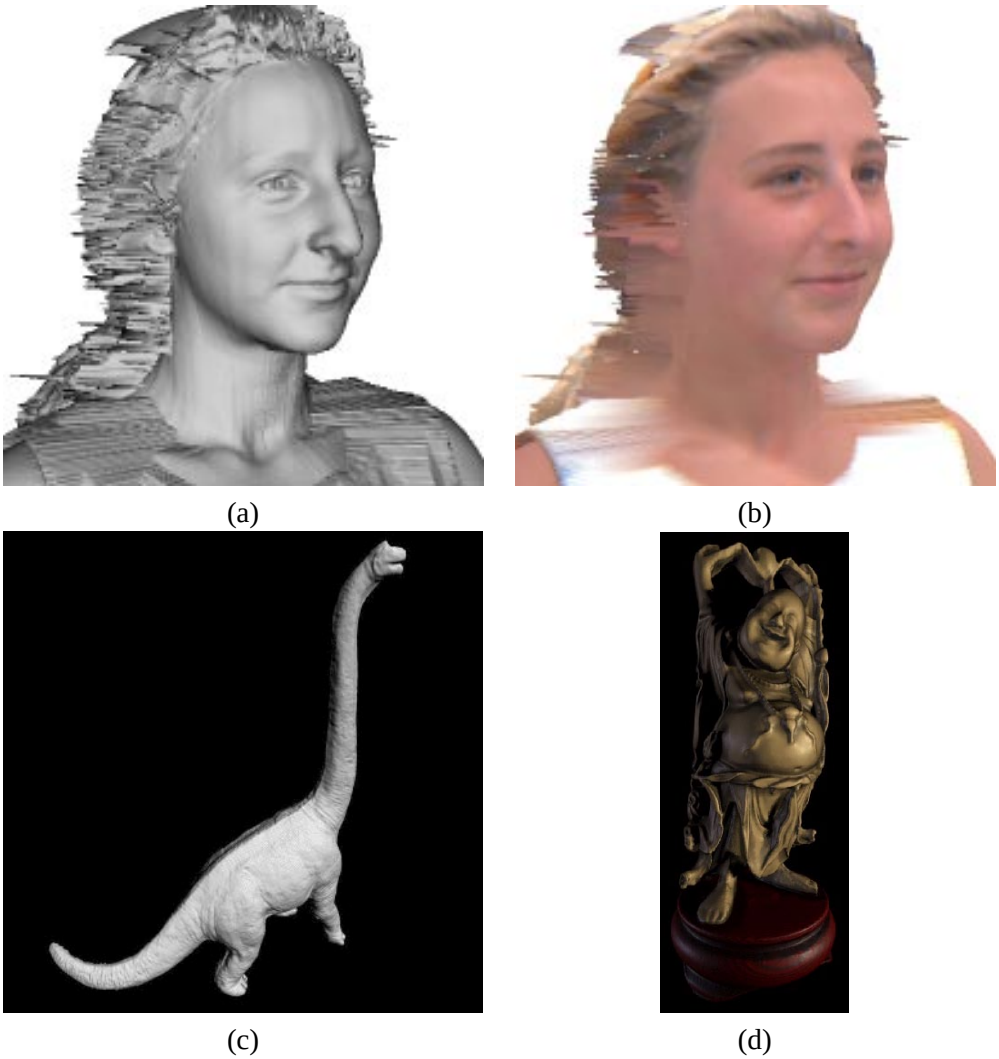
In the work presented in this thesis, we address this conundrum by showing that having an approximate model of the photographed scene can be used to robustly determine stereo correspondences from images taken from widely varying viewpoints. Specifically, the model enables us to warp the images to eliminate unequal foreshortening and to predict major instances of occlusion *before* trying to find correspondences. This new form of stereo is called *model-based stereo* and is presented in Chapter 7.

2.5 Range scanning

Instead of the anthropomorphic approach of using multiple images to reconstruct scene structure, an alternative technique is to use range imaging sensors [5] to directly measure depth to various points in the scene. Range imaging sensors determine depth either by triangulating the position of a projected laser stripe, or by measuring the time of flight of a directional laser pulse. Early versions of these sensors were slow, cumbersome and expensive. Although many improvements have been made, so far the most convincing demonstrations of the technology have been on human-scale objects and not on architectural scenes. Algorithms for combining multiple range images from different viewpoints have been developed both in computer vision [58, 42, 40] and in computer graphics [21, 53], see also Fig. 2.5. In many ways, range image based techniques and photographic techniques are complementary and have their relative advantages and disadvantages. Some advantages of modeling from photographic images are that (a) still cameras are inexpensive and widely available and (b) for some architecture that no longer exists all that is available are photographs. Furthermore, range images alone are insufficient for producing renderings of a scene; photometric information from photographs is also necessary.

2.6 Image-based modeling and rendering

In an image-based rendering system, the model consists of a set of images of a scene and their corresponding depth maps. When the depth of every point in an image is known, the image can be re-rendered from any nearby point of view by projecting the pixels of the image to their proper 3D locations and reprojecting them onto a new image plane. Thus, a new image of the scene is created by warping the images according to their depth maps. A principal attraction of image-based rendering is



(a) (b)

(c) (d)

Figure 2.5: Several models constructed from triangulation-based laser range scanning techniques. **(a)** A model of a person's head scanned using a commercially available Cyberware laser range scanner, using a cylindrical scan. **(b)** A texture-mapped version of this model, using imagery acquired by the same video camera used to detect the laser stripe. **(c)** A more complex geometry assembled by zippering together several triangle meshes obtained from separate linear range scans of a small object from [53]. **(d)** An even more complex geometry acquired from over sixty range scans using the volumetric recovery method in [7].

that it offers a method of rendering arbitrarily complex scenes with a constant amount of computation required per pixel. Using this property, [57] demonstrated how regularly spaced synthetic images (with their computed depth maps) could be warped and composited in real time to produce a virtual environment.

In [31], shown in Fig. 1.4, stereo photographs with a baseline of eight inches were taken every meter along a trail in a forest. Depth was extracted from each stereo pair using a census stereo algorithm [59]. Novel views were produced by supersampled z-buffered forward pixel splatting based on the stereo depth estimate of each pixel. ([24] describes a different rendering approach that implicitly triangulated the depth maps.) By manually determining relative camera pose between successive stereo pairs, it was possible to optically combine rerenderings from neighboring stereo pairs to fill in missing texture information. The project was able to produce very realistic synthetic views looking forward along the trail from any position within a meter of the original camera path, which was adequate for producing a realistic virtual experience of walking down the trail. Thus, for mostly linear environments such as a forest trail, this method of capture and rendering seems promising.

More recently, [27] presented a real-time image-based rendering system that used panoramic photographs with depth computed, in part, from stereo correspondence. One finding of the paper was that extracting reliable depth estimates from stereo is “very difficult”. The method was nonetheless able to obtain acceptable results for nearby views using user input to aid the stereo depth recovery: the correspondence map for each image pair was seeded with 100 to 500 user-supplied point correspondences and also post-processed. Even with user assistance, the images used still had to be closely spaced; the largest baseline described in the paper was five feet.

The requirement that samples be close together is a serious limitation to generating a freely

navigable virtual environment. Covering the size of just one city block would require thousands of panoramic images spaced five feet apart. Clearly, acquiring so many photographs is impractical. Moreover, even a dense lattice of ground-based photographs would only allow renderings to be generated from within a few feet of the original camera level, precluding any virtual fly-bys of the scene. Extending the dense lattice of photographs into three dimensions would clearly make the acquisition process even more difficult.

The modeling and rendering approach described in this thesis takes advantage of the structure in architectural scenes so that only a sparse set of photographs can be used to recover both the geometry and the appearance of an architectural scene. For example, our approach has yielded a virtual fly-around of a building from just twelve photographs (Fig. 5.14).

Some research done concurrently with the work presented in this thesis [4] also shows that taking advantage of architectural constraints can simplify image-based scene modeling. This work specifically explored the constraints associated with the cases of parallel and coplanar edge segments.

None of the work discussed so far has presented how to use intensity information coming from multiple photographs of a scene, taken from arbitrary locations, to render recovered geometry. The view-dependent texture mapping work (Chapter 6) presented in this thesis presents such a method.

Lastly, our model-based stereo algorithm (Chapter 7) presents an approach to robustly extracting detailed scene information from widely-spaced views by exploiting an approximate model of the scene.

Chapter 3

Overview

In this paper we present three new modeling and rendering techniques: photogrammetric modeling, view-dependent texture mapping, and model-based stereo. We show how these techniques can be used in conjunction to yield a convenient, accurate, and photorealistic method of modeling and rendering architecture from photographs. In our approach, the photogrammetric modeling program is used to create a basic volumetric model of the scene, which is then used to constrain stereo matching. Our rendering method composites information from multiple images with view-dependent texture-mapping. Our approach is successful because it splits the task of modeling from images into tasks which are easily accomplished by a person (but not a computer algorithm), and tasks which are easily performed by a computer algorithm (but not a person).

In Chapter 4, we discuss **camera calibration** from the standpoint of reconstructing architectural scenes from photographs. We present the method of camera calibration used in the work presented in this thesis, in which radial distortion is estimated separately from the perspective camera geometry. We also discuss methods of using uncalibrated views, which are quite often necessary

to work with when using historic photographs.

In Chapter 5, we present our **photogrammetric modeling** method. In essence, we have recast the structure from motion problem not as the recovery of individual point coordinates, but as the recovery of the parameters of a constrained hierarchy of parametric primitives. The result is that accurate architectural models can be recovered robustly from just a few photographs and with a minimal number of user-supplied correspondences.

In Chapter 6, we present **view-dependent texture mapping**, and show how it can be used to realistically render the recovered model. Unlike traditional texture-mapping, in which a single static image is used to color in each face of the model, view-dependent texture mapping interpolates between the available photographs of the scene depending on the user's point of view. This results in more lifelike animations that better capture surface specularities and unmodeled geometric detail.

In Chapter 7, we present **model-based stereo**, which is used to automatically refine a basic model of a photographed scene. This technique can be used to recover the structure of architectural ornamentation that would be difficult to recover with photogrammetric modeling. In particular, we show that projecting pairs of images onto an initial approximate model allows conventional stereo techniques to robustly recover very accurate depth measurements from images with widely varying viewpoints.

Lastly, in Chapter 8, we present ***Rouen Revisited***, an interactive art installation that represents an application of all the techniques developed in this thesis. This work, developed in collaboration with Interval Research Corporation, involved modeling the detailed Gothic architecture of the West façade of the Rouen Cathedral, and then rendering it from any angle, at any time of day, in any weather, and either as it stands today, as it stood one hundred years ago, or as the French

Impressionist Claude Monet might have painted it.

As we mentioned, our approach is successful not only because it synthesizes these geometry-based and image-based techniques, but because it divides the task of modeling from images into sub-tasks which are easily accomplished by a person (but not a computer algorithm), and sub-tasks which are easily performed by a computer algorithm (but not a person.) The correspondences for the reconstruction of the coarse model of the system are provided by the user in an interactive way; for this purpose we have designed and implemented *Façade* (Chapter 5), a photogrammetric modeling system that makes this task quick and simple. Our algorithm is designed so that the correspondences the user must provide are few in number per image. By design, the high-level model recovered by the system is precisely the sort of scene information that would be difficult for a computer algorithm to discover automatically. The geometric detail recovery is performed by an automated stereo correspondence algorithm (Chapter 7), which has been made feasible and robust by the pre-warping step provided by the coarse geometric model. In this case, corresponding points must be computed for a dense sampling of image pixels, a job too tedious to assign to a human, but feasible for a computer to perform using model-based stereo.