To appear in: Proc. 7th Int. Conf. on Computer Vision, Corfu, Greece, September 22-25, 1999

# A Theory of Shape by Space Carving

Kiriakos N. Kutulakos*
Depts. of Computer Science & Dermatology
University of Rochester
Rochester, NY 14627 USA

Steven M. Seitz†
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA

## Abstract

*In this paper we consider the problem of computing the 3D shape of an unknown, arbitrarily-shaped scene from multiple photographs taken at known but arbitrarily-distributed viewpoints. By studying the equivalence class of all 3D shapes that reproduce the input photographs, we prove the existence of a special member of this class, the* photo hull*, that (1) can be computed directly from photographs of the scene, and (2) subsumes all other members of this class. We then give a provably-correct algorithm, called* Space Carving*, for computing this shape and present experimental results on complex real-world scenes. The approach is designed to (1) build photorealistic shapes that accurately model scene appearance from a wide range of viewpoints, and (2) account for the complex interactions between occlusion, parallax, shading, and their effects on arbitrary views of a 3D scene.*

## 1. Introduction

A fundamental problem in computer vision is reconstructing the shape of a complex 3D scene from multiple photographs. While current techniques work well under controlled conditions (e.g., small stereo baselines [1], active viewpoint control [2], spatial and temporal smoothness [3], or scenes containing linear features or texture-less surfaces [4–6]), very little is known about scene reconstruction under general conditions. In particular, in the absence of *a priori* geometric information, what can we infer about the structure of an unknown scene from $N$ arbitrarily positioned cameras at known viewpoints? Answering this question has many implications for reconstructing real objects and environments, which tend to be non-smooth, exhibit significant occlusions, and may contain both textured and texture-less surface regions (Figure 1).

In this paper, we develop a theory for reconstructing arbitrarily-shaped scenes from arbitrarily-positioned cameras by formulating shape recovery as a constraint satisfaction problem. We show that any set of photographs of a rigid scene defines a collection of *picture constraints* that are satisfied by every scene projecting to those photographs. Furthermore, we characterize the set of all 3D shapes that satisfy these constraints and use the underlying theory to design a practical reconstruction algorithm, called *Space Carving*, that applies to fully-general shapes and camera configurations. In particular, we address three questions:

- Given $N$ input photographs, can we characterize the set of all *photo-consistent shapes*, i.e., shapes that reproduce the input photographs?
- Is it possible to compute a shape from this set and if so, what is the algorithm?
- What is the relationship of the computed shape to all other photo-consistent shapes?

Our goal is to study the $N$-view shape recovery problem in the general case where *no constraints* are placed upon the scene's shape or about the viewpoints of the input photographs. In particular, we address the above questions for the case when (1) no constraints are imposed on scene geometry or topology, (2) no constraints are imposed on the positions of the input cameras, (3) no information is available about the existence of specific image features in the input photographs (e.g., edges, points, lines, contours, texture, or color), and (4) no *a priori* correspondence information is available. Unfortunately, even though several algorithms have been proposed for recovering shape from multiple views that work under some of these conditions (e.g., work on stereo [7–9]), very little is currently known about how to answer the above questions, and even less so about how to answer them in this general case.

At the heart of our work is the observation that these questions become tractable when scene radiance belongs to a general class of radiance functions we call *locally computable*. This class characterizes scenes for which global illumination effects such as shadows, transparency and interreflections can be ignored, and is sufficiently general to include scenes with parameterized radiance models (e.g., Lambertian, Phong, Torrance-Sparrow [10]). Using this observation as a starting point, we show how to compute,

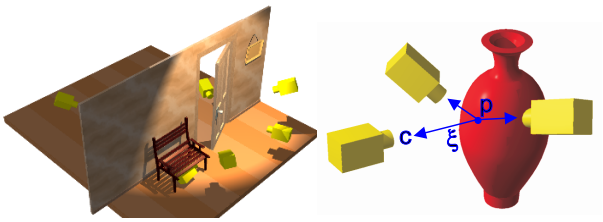
* Kiriakos Kutulakos gratefully acknowledges the support of the National Science Foundation under Grant No. IRI-9875628, of Roche Laboratories, Inc., and of the Dermatology Foundation.
† Part of this work was conducted while Steven Seitz was employed by the Vision Technology Group at Microsoft Research. The support of the Microsoft Corporation is gratefully acknowledged.

from $N$ photographs of an unknown scene, a maximal shape called the *photo hull* that encloses the set of all photo-consistent reconstructions. The only requirements are that (1) the viewpoint of each photograph is known in a common 3D world reference frame (Euclidean, affine, or projective, and (2) scene radiance follows a known, locally-computable radiance function. Experimental results illustrating our method's performance are given for both real and simulated geometrically-complex scenes.

To our knowledge, no previous theoretical work has studied the equivalence class of solutions to the general $N$-view reconstruction problem or provably-correct algorithms for computing them.[1] The Space Carving Algorithm that results from our analysis, however, is related to other 3D scene-space stereo algorithms that have been recently proposed [14–21]. Of these, most closely related are mesh-based [14] and level-set [22] algorithms, as well as methods that sweep a plane or other manifold through a discretized scene space [15–17, 20, 23]. While the algorithms in [14, 22] generate high-quality reconstructions and perform well in the presence of occlusions, their use of regularization techniques penalizes complex surfaces and shapes. Even more importantly, no formal study has been undertaken to establish their validity for recovering arbitrarily-shaped scenes from unconstrained camera configurations (e.g., the one shown in Figure 1a). In contrast, our Space Carving Algorithm is provably correct and has no regularization biases. Even though space-sweep approaches have many attractive properties, existing algorithms [15–17, 20] are not fully general i.e., they rely on the presence of specific image features such as edges and hence generate only sparse reconstructions [15], or they place strong constraints on the input viewpoints relative to the scene [16, 17]. Unlike all previous methods, Space Carving guarantees complete reconstruction in the general case.

Our approach offers four main contributions over the existing state of the art. First, it introduces an algorithm-independent analysis of the $N$ view shape-recovery problem, making explicit the assumptions required for solving it as well as the ambiguities intrinsic to the problem. Second, it establishes the tightest possible bound on the shape of the true scene obtainable from $N$ photographs without *a priori* geometric information. Third, it describes the first provably-correct algorithm for scene reconstruction from unconstrained camera viewpoints. Fourth, the approach leads naturally to global reconstruction algorithms that recover 3D shape information from all photographs at once, eliminating the need for complex partial reconstruction and merging operations [19, 24].

[1] Faugeras [11] has recently proposed the term *metameric* to describe such shapes, in analogy with the term's use in the color perception [12] and structure-from-motion literature [13].




(a) (b)

**Figure 1.** Viewing geometry. The scene volume and camera distribution covered by our analysis are both completely unconstrained. Examples include (a) a 3D environment viewed from a collection of cameras that are arbitrarily dispersed in free space, and (b) a 3D object viewed by a single camera moving around it.

## 2. Picture Constraints

Let $\mathcal{V}$ be a 3D scene defined by a finite, opaque, and possibly disconnected volume in space. We assume that $\mathcal{V}$ is viewed under perspective projection from $N$ known positions $c_1, \ldots, c_N$ in $\Re^3 - \mathcal{V}$ (Figure 1b). The *radiance* of a point $p$ on the scene's surface is a function $rad_p(\xi)$ that maps every oriented ray $\xi$ through the point to the color of light reflected from $p$ along $\xi$. We use the term *shape-radiance scene description* to denote the scene $\mathcal{V}$ together with an assignment of a radiance function to every point on its surface. This description contains all the information needed to reproduce a photograph of the scene for any camera position.

Every photograph of a 3D scene taken from a known location partitions the set of all possible shape-radiance scene descriptions into two families, those that reproduce the photograph and those that do not. We characterize this constraint for a given shape and a given radiance assignment by the notion of *photo-consistency*:[2]

**Definition 1 (Point Photo-Consistency)** *A point $p$ in $\mathcal{V}$ that is visible from $c$ is* photo-consistent with the photograph at $c$ *if (1) $p$ does not project to a background pixel, and (2) the color at $p$'s projection is equal to* $\mathrm{rad}_p(\vec{pc})$.

**Definition 2 (Shape-Radiance Photo-Consistency)** *A shape-radiance scene description is* photo-consistent with the photograph at $c$ *if all points visible from $c$ are photo-consistent and every non-background pixel is the projection of a point in $\mathcal{V}$.*

**Definition 3 (Shape Photo-Consistency)** *A shape $\mathcal{V}$ is* photo-consistent *with a set of photographs if there is an assignment of radiance functions to the visible points of $\mathcal{V}$ that makes the resulting shape-radiance description photo-consistent with all photographs.*

[2] In the following, we make the simplifying assumption that pixel values in the image measure scene radiance directly.

Our goal is to provide a concrete characterization of the family of all scenes that are photo-consistent with $N$ input photographs. We achieve this by making explicit the two ways in which photo-consistency with $N$ photographs can constrain a scene's shape.

## 2.1. Background Constraints

Photo-consistency requires that no point of $\mathcal{V}$ projects to a background pixel. If a photograph taken at position $c$ contains identifiable background pixels, this constraint restricts $\mathcal{V}$ to a cone defined by $c$ and the photograph's non-background pixels. Given $N$ such photographs, the scene is restricted to the *visual hull*, which is the volume of intersection of their corresponding cones [5].

When no *a priori* information is available about the scene's radiance, the visual hull defines all the shape constraints in the input photographs. This is because there is always an assignment of radiance functions to the points on the surface of the visual hull that makes the resulting shape-radiance description photo-consistent with the $N$ input photographs.[3] The visual hull can therefore be thought of as a "least commitment reconstruction" of the scene—any further refinement of this volume must rely on assumptions about the scene's shape or radiance.

While visual hull reconstruction has often been used as a method for recovering 3D shape from photographs [25, 26], the picture constraints captured by the visual hull only exploit information from the background pixels in these photographs. Unfortunately, these constraints become useless when photographs contain no background pixels (i.e., the visual hull degenerates to $\Re^3$) or when background identification cannot be performed accurately. Below we study picture constraints from non-background pixels when the scene's radiance is restricted to a special class of radiance models. The resulting constraints lead to photo-consistent scene reconstructions that are subsets of the visual hull, and unlike the visual hull, can contain concavities.

## 2.2. Radiance Constraints

Surfaces that are not transparent or mirror-like reflect light in a coherent manner, i.e., the color of light reflected from a single point along different directions is not arbitrary. This coherence provides additional picture constraints beyond what can be obtained from background information. In order to take advantage of these constraints, we focus on scenes whose radiance satisfies the following criterion:

> **Consistency Check Criterion:** An algorithm $\texttt{consist}_K()$ is available that takes as input at least $K \leq N$ colors $col_1, \ldots, col_K$, $K$ vectors $\xi_1, \ldots, \xi_K$, and the light source positions (non-Lambertian case), and decides whether it is possible for a single surface point to reflect light of color $col_i$ in direction $\xi_i$ simultaneously for all $i = 1, \ldots, K$.

[3] For example, set $rad_p(\vec{pc})$ equal to the color at $p$'s projection.

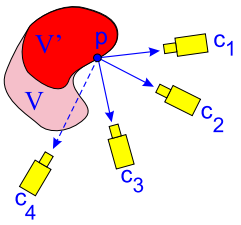



**Figure 2.** (a) Illustration of the Visibility and Non-Photo-Consistency Lemmas. If $p$ is non-photo-consistent with the photographs at $c_1, c_2, c_3$, it is non-photo-consistent with the entire set $Vis_{\mathcal{V}'}(p)$, which also includes $c_4$.

Given a shape $\mathcal{V}$, the Consistency Check Criterion gives us a way to establish the photo-consistency of every point on $\mathcal{V}$'s surface. This criterion defines a general class of radiance models, that we call *locally computable*, that are characterized by a locality property: the radiance at any point is independent of the radiance of all other points in the scene. The class of locally-computable radiance models therefore restricts our analysis to scenes where global illumination effects such as transparency, inter-reflection, and shadows can be ignored. This class subsumes the Lambertian ($K = 2$) and other parameterized radiance models.[4]

Given an *a priori* locally computable radiance model for the scene, we can determine whether or not a given shape $\mathcal{V}$ is photo-consistent with a collection of photographs. Even more importantly, when the scene's radiance is described by such a model, the *non*-photo-consistency of a shape $\mathcal{V}$ tells us a great deal about the shape of the underlying scene. We use the following two lemmas to make explicit the structure of the family of photo-consistent shapes. These lemmas provide the analytical tools needed to describe how the non-photo-consistency of a shape $\mathcal{V}$ affects the photo-consistency of its subsets (Figure 2):

**Lemma 1 (Visibility Lemma)** *Let $p$ be a point on $\mathcal{V}$'s surface,* $\text{Surf}(\mathcal{V})$*, and let* $\text{Vis}_{\mathcal{V}}(p)$ *be the collection of input photographs in which $\mathcal{V}$ does not occlude $p$. If $\mathcal{V}' \subset \mathcal{V}$ is a shape that also has $p$ on its surface,* $\text{Vis}_{\mathcal{V}}(p) \subseteq \text{Vis}_{\mathcal{V}'}(p)$.

*Proof:* Since $\mathcal{V}'$ is a subset of $\mathcal{V}$, no point of $\mathcal{V}'$ can lie between $p$ and the cameras corresponding to $Vis_{\mathcal{V}}(p)$. *QED*

**Lemma 2 (Non-Photo-Consistency Lemma)** *If* $p \in \text{Surf}(\mathcal{V})$ *is not photo-consistent with a subset of* $\text{Vis}_{\mathcal{V}}(p)$*, it is not photo-consistent with* $\text{Vis}_{\mathcal{V}}(p)$.

[4] Specific examples include (1) using a mobile camera mounted with a light source to capture photographs of a scene whose reflectance can be expressed in closed form (e.g., using the Torrance-Sparrow model [10, 27]), and (2) using multiple cameras to capture photographs of an approximately Lambertian scene under arbitrary unknown illumination (Figure 1).

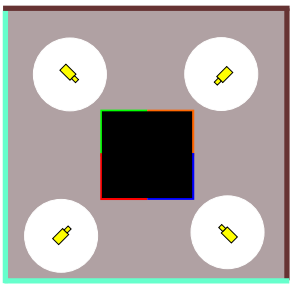


**Figure 3.** Trivial shape solutions in the absence of free-space constraints. A two-dimensional object consisting of a black square whose sides are painted four distinct diffuse colors (red, blue, orange, and green), is viewed by four cameras. Carving out a small circle around each camera and projecting the image onto the interior of that circle yields a trivial photo-consistent shape.

Intuitively, Lemmas 1 and 2 suggest that both visibility and non-photo-consistency exhibit a form of "monotonicity:" the Visibility Lemma tells us that the collection of photographs from which a surface point $p \in \mathit{Surf}(\mathcal{V})$ is visible strictly expands as $\mathcal{V}$ gets smaller (Figure 2). Analogously, the Non-Photo-Consistency Lemma, which follows as a direct consequence of the definition of photo-consistency, tells us that each new photograph can be thought of as an additional constraint on the photo-consistency of surface points—the more photographs are available, the more difficult it is for those points to achieve photo-consistency. Furthermore, once a surface point fails to be photo-consistent no new photograph of that point can re-establish photo-consistency.

The key consequence of Lemmas 1 and 2 is given by the following theorem which shows that *non*-photo-consistency at a point rules out the photo-consistency of an entire family of shapes:

**Theorem 1 (Subset Theorem)** *If $p \in \mathrm{Surf}(\mathcal{V})$ is not photo-consistent, no photo-consistent subset of $\mathcal{V}$ contains $p$.*

*Proof:* Let $\mathcal{V}' \subset \mathcal{V}$ be a shape that contains $p$. Since $p$ lies on the surface of $\mathcal{V}$, it must also lie on the surface of $\mathcal{V}'$. From the Visibility Lemma it follows that $\mathit{Vis}_{\mathcal{V}}(p) \subseteq \mathit{Vis}_{\mathcal{V}'}(p)$. The theorem now follows by applying the Non-Photo-Consistency Lemma to $\mathcal{V}'$ and using the locality property of locally computable radiance models. *QED*

We explore the ramifications of the Subset Theorem in the next section.

## 3. The Photo Hull

The family of all shapes that are photo-consistent with $N$ photographs defines the ambiguity inherent in the problem of recovering 3D shape from those photographs. When there is more than one photo-consistent shape it is impossible to decide, based on those photographs alone, which photo-consistent shape corresponds to the true scene. This ambiguity raises two important questions regarding the feasibility of scene reconstruction from photographs:

- Is it possible to compute a shape that is photo-consistent with $N$ photographs and, if so, what is the algorithm?
- If a photo-consistent shape can be computed, how can we relate that shape to all other photo-consistent 3D interpretations of the scene?

Before providing a general answer to these questions we observe that when the number of input photographs is finite, the first question can be answered with a trivial shape (Figure 3). In general, trivial shape solutions such as this one can only be eliminated with the incorporation of *free space* constraints, i.e., regions of space that are known not to contain scene points. Our analysis enables the (optional) inclusion of such constraints by specifying an arbitrary shape $\mathcal{V}$ within which a photo-consistent scene is known to lie.[5]

In particular, our answers to both questions rest on the following theorem. Theorem 2 shows that for any shape $\mathcal{V}$ there is a unique photo-consistent shape that subsumes, i.e., contains within its volume, all other photo-consistent shapes in $\mathcal{V}$ (Figure 4):

**Theorem 2 (Photo Hull Theorem)** *Let $\mathcal{V}$ be an arbitrary set of points and let $\mathcal{V}^*$ be the union of all photo-consistent subsets of $\mathcal{V}$. The shape $\mathcal{V}^*$ is photo-consistent and is called the* photo hull.[6]

*Proof:* (By contradiction) Suppose $\mathcal{V}^*$ is not photo-consistent and let $p$ be a non-photo-consistent point on its surface. Since $p \in \mathcal{V}^*$, there exists a photo-consistent shape, $\mathcal{V}' \subset \mathcal{V}^*$, that also has $p$ on its surface. It follows from the Subset Theorem that $\mathcal{V}'$ is not photo-consistent. *QED*

Theorem 2 provides an explicit relation between the photo hull and all other possible 3D interpretations of the scene: the theorem guarantees that every such interpretation is a subset of the photo hull. The photo hull therefore represents a least-commitment reconstruction of the scene. We describe a volumetric algorithm for computing this shape in the next section.

## 4. Reconstruction by Space Carving

An important feature of the photo hull is that it can be computed using a simple, discrete algorithm that "carves" space in a well-defined manner. Given an initial volume

[5] Note that if $\mathcal{V} = \Re^3$, the problem reduces to the case when no constraints on free space are available.

[6] Our use of the term *photo hull* to denote the "maximal" photo-consistent shape defined by a collection of photographs is due to a suggestion by Leonard McMillan.

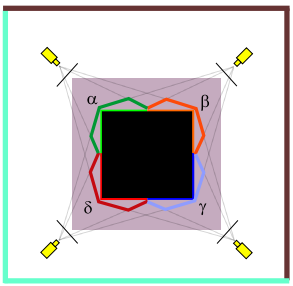



**Figure 4.** Illustration of the Photo Hull Theorem. The gray-shaded region corresponds to an arbitrary shape $\mathcal{V}$ containing the object of Figure 3. $\mathcal{V}^*$ is a polygonal region that extends beyond the true scene and whose boundary is defined by the polygonal segments $\alpha, \beta, \gamma$, and $\delta$. When these segments are colored as shown, $\mathcal{V}^*$'s projections are indistinguishable from that of the true object and *no* photo-consistent shape in the gray-shaded region can contain points outside $\mathcal{V}^*$.

$\mathcal{V}$ that contains the scene, the algorithm proceeds by iteratively removing (i.e. "carving") portions of that volume until it converges to the photo hull, $\mathcal{V}^*$. The algorithm can therefore be fully specified by answering four questions: (1) how do we select the initial volume $\mathcal{V}$, (2) how should we represent that volume to facilitate carving, (3) how do we carve at each iteration to guarantee convergence to the photo hull, and (4) when do we terminate carving?

The choice of the initial volume has a considerable impact on the outcome of the reconstruction process (Figure 3). Nevertheless, selection of this volume is beyond the scope of this paper; it will depend on the specific 3D shape recovery application and on information about the manner in which the input photographs were acquired.[7] Below we consider a general algorithm that, given $N$ photographs and *any* initial volume that contains the scene, is guaranteed to find the (unique) photo hull contained in that volume.

In particular, let $\mathcal{V}$ be an arbitrary finite volume that contains the scene as an unknown sub-volume. Also, assume that the surface of the true scene conforms to a radiance model defined by a consistency check algorithm $\texttt{consist}_K()$. We represent $\mathcal{V}$ as a finite collection of voxels $v_1, \ldots, v_M$. Using this representation, each carving iteration removes a single voxel from $\mathcal{V}$.

The Subset Theorem leads directly to a method for selecting a voxel to carve away from $\mathcal{V}$ at each iteration. Specifically, the theorem tells us that if a voxel $v$ on the surface of $\mathcal{V}$ is not photo-consistent, the volume $\mathcal{V} = \mathcal{V} - \{v\}$ must still contain the photo hull. Hence, if only non-photo-consistent voxels are removed at each iteration, the carved volume is guaranteed to converge to the photo hull.

[7] Examples include defining $\mathcal{V}$ to be equal to the visual hull or, in the case of a camera moving through an environment , $\Re^3$ minus a tube along the camera's path.

The order in which non-photo-consistent voxels are examined and removed is not important for guaranteeing correctness. Convergence to this shape occurs when no non-photo-consistent voxel can be found on the surface of the carved volume. These considerations lead to the following algorithm for computing the photo hull:[8]

### Space Carving Algorithm

**Step 1:** Initialize $\mathcal{V}$ to a volume containing the true scene.

**Step 2:** Repeat the following steps for voxels $v \in \mathit{Surf}(\mathcal{V})$ until a non-photo-consistent voxel is found:

- **a.** Project $v$ to all photographs in $\mathit{Vis}_{\mathcal{V}}(v)$. Let $col_1, \ldots, col_j$ be the pixel colors to which $v$ projects and let $\xi_1, \ldots, \xi_j$ be the optical rays connecting $v$ to the corresponding optical centers.
- **b.** Determine the photo-consistency of $v$ using $\texttt{consist}_K(col_1, \ldots, col_j, \xi_1, \ldots, \xi_j)$.

**Step 3:** If no non-photo-consistent voxel is found, set $\mathcal{V}^* = \mathcal{V}$ and terminate. Otherwise, set $\mathcal{V} = \mathcal{V} - \{v\}$ and repeat Step 2.

The key step in the algorithm is the search and voxel consistency checking of Step 2. The following proposition gives an upper bound on the number of voxel photo-consistency checks:[9]

**Proposition 1** *The total number of required photo-consistency checks is bounded by $N * M$ where $N$ is the number of input photographs and $M$ is the number of voxels in the initial (i.e., uncarved) volume.*

To perform visibility computations efficiently, we use a multi-sweep implementation of space carving. In what follows, we briefly summarize the technique, but full details of the approach are omitted due to space limitations. Each pass consists of sweeping a plane through the scene volume and testing the photo-consistency of voxels on that plane. The advantage of this method is that voxels are always visited in an order that captures all occlusion relations between the entire set of voxels and an appropriately-chosen subset $\mathcal{C}$ of the cameras: each sweep guarantees that if a voxel $p$ occludes another voxel $q$ when viewed from a camera in $\mathcal{C}$, $p$ will necessarily be visited before $q$. This is achieved by choosing $\mathcal{C}$ to be the set of all cameras that lie on one side of the sweep plane. Since each plane sweep considers only a subset of the cameras from which a voxel may be visible, multiple sweeps are needed to ensure photo-consistency of voxels with *all* input views. Our implementation cycles through six directions in each pass, i.e., in increasing/decreasing $x$, $y$, and $z$ directions, and applies repeated passes until the carving procedure converges. In practice, this typically occurs after 2 or 3 passes.

[8] Convergence to this shape is provably guaranteed only for scenes representable by a discrete set of voxels.

[9] Proof is omitted due to lack of space.

## 5. Experimental Results

To demonstrate the applicability of our approach, we performed several experiments on real and synthetic image sequences. In all examples, a Lambertian model was used for the Consistency Check Criterion, i.e., it was assumed that a voxel projects to pixels of approximately the same color in every image. We used a threshold on the standard deviation of these pixels to decide whether or not to carve a voxel.

We first ran the Space Carving Algorithm on 16 images of a gargoyle sculpture (Figs. 5a-e). The sub-pixel calibration error in this sequence enabled using a small threshold of 6% for the RGB component error. This threshold, along with the voxel size and the 3D coordinates of a bounding box containing the object were the only parameters given as input to our implementation. Some errors are still present in the reconstruction, notably holes that occur as a result of shadows and other illumination changes caused by moving the object rather than camera. These effects were not accounted for by the radiance model, causing some voxels to be erroneously carved. The finite voxel size, calibration error, and image discretization effects resulted in a loss of some fine surface detail. Voxel size could be further reduced with better calibration, but only up to the point where image discretization effects (i.e., finite pixel size) become a significant source of error. Figs. 5f-i show results from applying our algorithm to images of a human hand.

In a final experiment, we applied our algorithm to images of a synthetic building scene rendered from both its interior and exterior (Figure 6). This placement of cameras yields an extremely difficult stereo problem, due to the drastic changes in visibility between interior and exterior cameras.[10] Figure 6 compares the original model and the reconstruction from different viewpoints. The model's appearance is very good near the input viewpoints, as demonstrated in Figs. 6b-c. Note that the reconstruction tends to "bulge" out and that the walls are not perfectly planar (Figure 6e). This behavior is exactly as predicted by Theorem 2—the algorithm converges to the *largest possible* shape that is consistent with the input images. In low-contrast regions where shape is visually ambiguous, this causes significant deviations between the computed photo hull and the true scene. While these deviations do not adversely affect scene appearance near the input viewpoints, they can result in noticeable artifacts for far-away views. These deviations and the visual artifacts they cause are easily remedied by including images from a wider range of camera viewpoints to further constrain the scene's shape, as shown in Figure 6f.

Our experiments highlight a number of advantages of our approach over previous techniques. Existing multi-baseline stereo techniques [1] work best for densely textured scenes and suffer in the presence of large occlusions. In contrast, the gargoyle and hand sequences contain many low-textured regions and dramatic changes in visibility. While contour-based techniques like volume intersection [25] work well for similar scenes, they require detecting silhouettes or occluding contours. For the gargoyle sequence, the background was unknown and heterogeneous, making the contour detection problem extremely difficult. Note also that Seitz and Dyer's voxel coloring technique [16] would not work for any of the above sequences because of the constraints it imposes on camera placement. Our approach succeeds because it integrates both texture and contour information as appropriate, without the need to explicitly detect features or contours, or constrain viewpoints. Our results indicate the approach is highly effective for both densely textured and untextured objects and scenes.

[10] For example, the algorithms in [16, 17] fail catastrophically for this scene because the distribution of the input views and the resulting occlusion relationships violate the assumptions used by those algorithms.

## 6. Concluding Remarks

While the Space Carving Algorithm's effectiveness was demonstrated in the presence of image noise, the photo-consistency theory itself is based on an idealized model of image formation. Extending the theory to explicitly model image noise, quantization and calibration errors, and their effects on the photo hull is an open research problem. Extending the formulation to handle non-locally computable radiance models (e.g., shadows) is another important topic of future work. Other research directions include (1) developing space carving algorithms for noisy images, (2) investigating the use of surface-based rather than voxel-based techniques for finding the photo hull, (3) incorporating *a priori* shape constraints (e.g., smoothness), and (4) analyzing the topological structure of the set of photo-consistent shapes.

## References


[1] M. Okutomi and T. Kanade, "A multiple-baseline stereo," *T-PAMI*, v.15, pp. 353–363, 1993.

[2] K. N. Kutulakos and C. R. Dyer, "Recovering shape by purposive viewpoint adjustment," *IJCV*, v.12, n.2, pp. 113–136, 1994.

[3] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *IJCV*, v.1, pp. 7–55, 1987.

[4] R. Cipolla and A. Blake, "Surface shape from the deformation of apparent contours," *IJCV*, v.9, n.2, pp. 83–112, 1992.

[5] A. Laurentini, "The visual hull concept for silhouette-based image understanding," *T-PAMI*, v.16, pp. 150–162, 1994.

[6] K. N. Kutulakos and C. R. Dyer, "Global surface reconstruction by purposive control of observer motion," *Artificial Intelligence Journal*, v.78, n.1-2, pp. 147–177, 1995.

[7] P. N. Belhumeur, "A bayesian approach to binocular stereopsis," *IJCV*, v.19, n.3, pp. 237–260, 1996.

[8] I. Cox, S. Hingorani, S. Rao, and B. Maggs, "A maximum likelihood stereo algorithm," *CVIU: Image Understanding*, v.63, n.3, pp. 542–567, 1996.

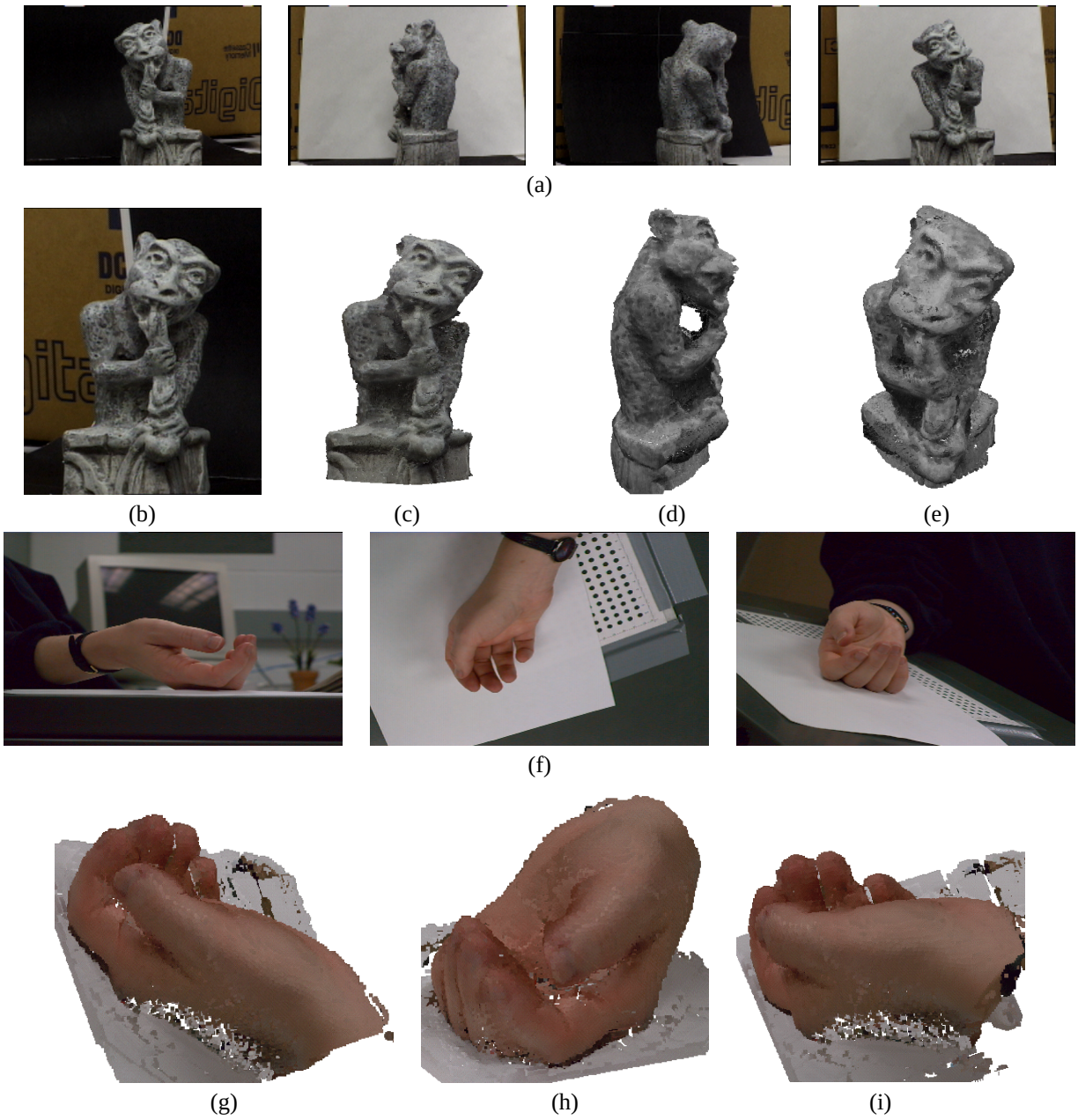



**Figure 5.** Reconstruction results for two real scenes. (a-e) Reconstruction of a gargoyle stone sculpture. Four out of 16 486x720 RGB input images are shown in (a). The images were acquired by (1) rotating the object in front of a stationary background in $22.5^\circ$ increments, and (2) altering the object's background before each image was acquired. This latter step enabled complete reconstruction of the sculpture without any initial segmentation step—the space carving process ensured that photo-consistency could not be enforced for points projecting to non-object pixels. (b-e) Reconstruction of the sculpture. The model contains 215,000 surface voxels, carved out of an initial volume of approximately 51 million. It took 250 minutes to compute on an SGI O2 R10000/175MHz workstation. One of the 16 input images is shown (b), along with views of the reconstruction from the same (c) and new (d-e) viewpoints. (f-i) Reconstruction of a hand. Three out of one hundred calibrated images of the hand are shown in (f). The sequence was generated by moving a camera around the subject's stationary hand. (g-i) Views of the reconstruction. The reconstructed model was computed using an RGB component error threshold of 15. The model has 112,000 voxels and took 53 minutes to compute on the SGI O2 using a graphics hardware-accelerated version of the algorithm.

[9] C. V. Stewart, "MINPRAN: A new robust estimator for computer vision," *T-PAMI.*, v.17, pp. 925–938, 1995.

[10] K. E. Torrance and E. M. Sparrow, "Theory of off-specular reflection from roughened surface," *JOSA*, v.57, pp.1105–1114, 1967.

[11] O. D. Faugeras, "Personal communication."

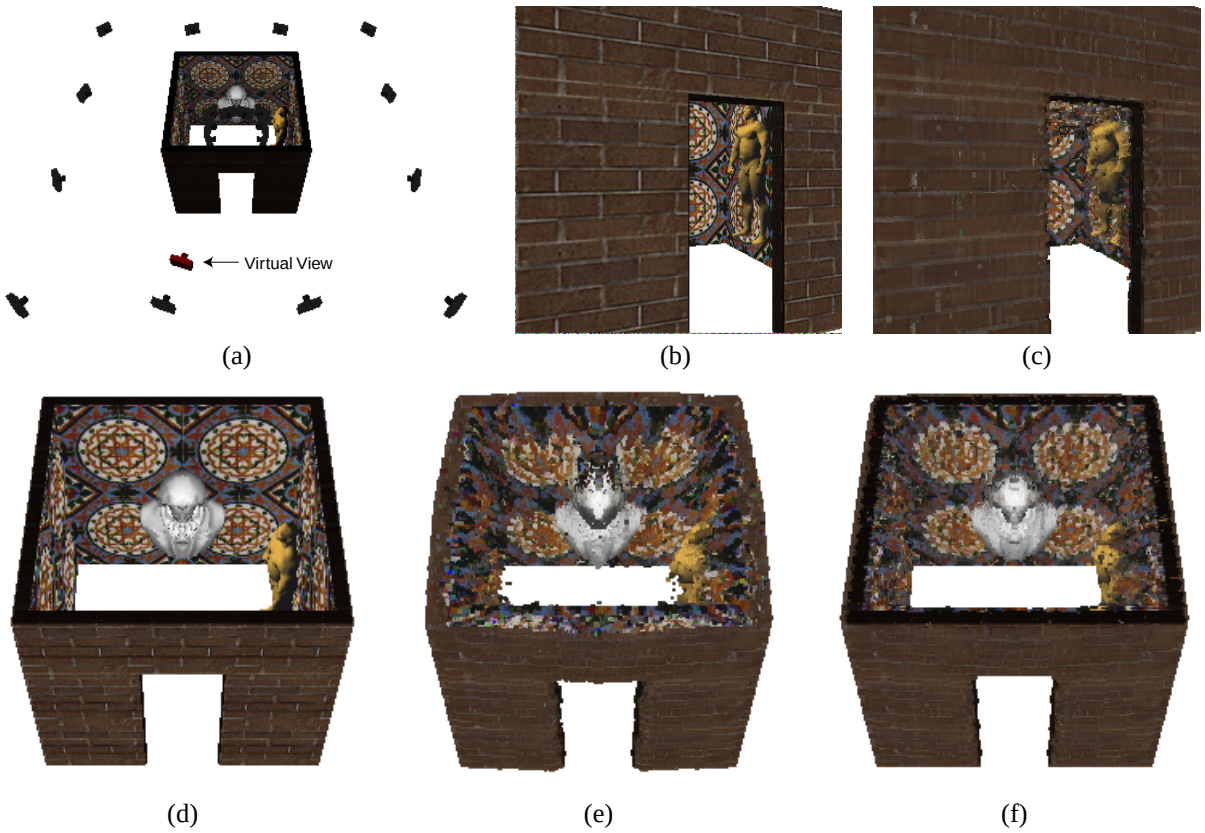



**Figure 6.** Reconstruction of a synthetic building scene. (a) 24 Cameras were placed in both the interior and exterior of a building to enable simultaneous, complete reconstruction of its exterior and interior surfaces. The reconstruction contains 370,000 voxels, carved out of a $200 \times 170 \times 200$ voxel block. (b) A rendered image of the building for a viewpoint near the input cameras (shown as "virtual view" in (a)) is compared to the view of the reconstruction (c). (d-f) Views of the reconstruction from far away camera viewpoints. (d) shows a rendered top view of the original building, (e) the same view of the reconstruction, and (f) a new reconstruction resulting from adding image (d) to the set of input views. Note that adding just a single top view dramatically improves the quality of the reconstruction.


[12] R. L. Alfvin and M. D. Fairchild, "Observer variability in metameric color matches using color reproduction media," *Color Research & Application*, v.22, n.3, pp. 174–178, 1997.

[13] J. A. J. C. van Veen and P. Werkhoven, "Metamerisms in structure-from-motion perception," *Vision Research*, v.36, pp. 2197–2210, 1996.

[14] P. Fua and Y. G. Leclerc, "Object-centered surface reconstruction: Combining multi-image stereo and shading," *IJCV*, v.16, pp. 35–56, 1995.

[15] R. T. Collins, "A space-sweep approach to true multi-image matching," *Proc. CVPR*, pp. 358–363, 1996.

[16] S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by voxel coloring," *Proc. CVPR*, pp. 1067–1073, 1997.

[17] S. M. Seitz and K. N. Kutulakos, "Plenoptic image editing," *Proc. ICCV*, pp. 17–24, 1998.

[18] C. L. Zitnick and J. A. Webb, "Multi-baseline stereo using surface extraction," Tech. Rep. CMU-CS-96-196, Carnegie Mellon University, Pittsburgh, PA, November 1996.

[19] P. J. Narayanan, P. W. Rander, and T. Kanade, "Constructing virtual worlds using dense stereo," *Proc. ICCV*, pp. 3–10, 1998.

[20] R. Szeliski and P. Golland, "Stereo matching with transparency and matting," *Proc. ICCV*, pp. 517–524, 1998.

[21] S. Roy and I. J. Cox, "A maximum-flow formulation of the N-camera stereo correspondence problem," *Proc. ICCV*, pp. 492–499, 1998.

[22] O. Faugeras and R. Keriven, "Complete dense stereovision using level set methods," *Proc. ECCV*, pp. 379–393, 1998.

[23] M. S. Langer and S. W. Zucker, "Shape-from-shading on a cloudy day," *JOSA-A*, v.11, n.2, pp. 467–478, 1994.

[24] W. B. Seales and O. Faugeras, "Building three-dimensional object models from image sequences," *CVIU*, v.61, n.3, pp. 308–324, 1995.

[25] R. Szeliski, "Rapid octree construction from image sequences," *CVGIP: Image Understanding*, v.58, n.1, pp. 23–32, 1993.

[26] K. N. Kutulakos, "Shape from the light field boundary," *Proc. CVPR*, pp. 53–59, 1997.

[27] Y. Sato, M. D. Wheeler, and K. Ikeuchi, "Object shape and reflectance modeling from observation," *Proc. SIGGRAPH*, pp. 379–387, 1997.