International Journal of Computer Vision 38(3), 199–218, 2000

A Theory of Shape by Space Carving

KIRIAKOS N. KUTULAKOS
Department of Computer Science and Department of Dermatology, University of Rochester, Rochester, NY 14627, USA
kyros@cs.rochester.edu

STEVEN M. SEITZ
The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
seitz@cs.cmu.edu

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) capture 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 view-dependent effects on scene-appearance.

Keywords: scene modeling, photorealistic reconstruction, multi-view stereo, space carving, voxel coloring, shape-from-silhouettes, visual hull, volumetric shape representations, metameric shapes, 3D photography

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 (Okutomi and Kanade, 1993), active viewpoint control (Kutulakos and Dyer, 1994), spatial and temporal smoothness (Poggio et al., 1985; Bolles et al., 1987; Katayama et al., 1995), or scenes containing curved lines (Bascle and Deriche, 1993), planes (Pritchett and Zisserman, 1998), or texture-less surfaces (Cipolla and Blake, 1992; Vaillant and Faugeras, 1992; Laurentini, 1994; Szeliski and Weiss, 1994; Kutulakos and Dyer, 1995)), 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 (Fig. 1).

In this paper, we develop a theory for reconstructing 3D scenes from photographs 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?

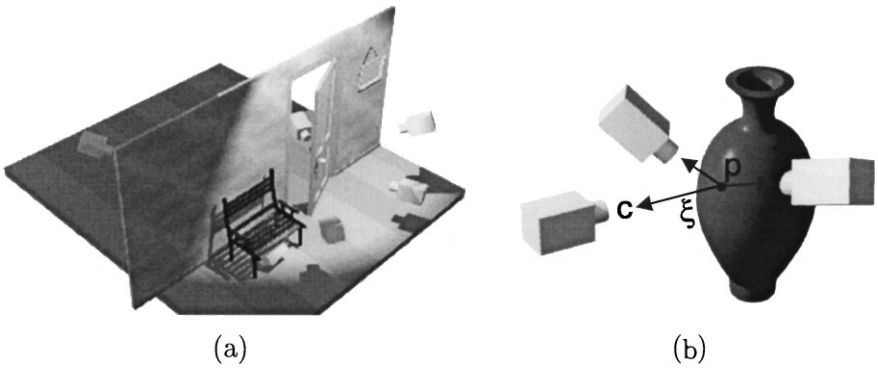


Figure 1. Viewing geometry. The scene volume and camera distribution covered by our analysis are both completely unconstrained. Examples include (a) a 3D environment viewed from a collection of cameras that are arbitrarily dispersed in free space, and (b) a 3D object viewed by a single camera moving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

– 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 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 (Belhumeur, 1996; Cox et al., 1996; Stewart, 1995)), 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 inter-reflections can be ignored, and is sufficiently general to include scenes with parameterized radiance models (e.g., Lambertian, Phong (Foley et al., 1990), Torrance-Sparrow (Torrance and Sparrow, 1967)). Using this observation as a starting point, we show how to compute, 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 (Koenderink and van Doorn, 1991), or projective (Mundy and Zisserman, 1992)), and (2) scene radiance follows a known, locally-computable radiance function. Experimental results demonstrating our method's performance are given for both real and simulated geometrically-complex scenes.

Central to our analysis is the realization that parallax, occlusion, and scene radiance all contribute to a photograph's dependence on viewpoint. Since our notion of photo-consistency implicitly ensures that all of these 3D shape cues are taken into account in the recovery process, our approach is related to work on stereo (Okutomi and Kanade, 1993; Cox et al., 1996; Hoff and Ahuja, 1989), shape-from-contour (Cipolla and Blake, 1992; Vaillant and Faugeras, 1992; Szeliski, 1993), as well as shape-from-shading (Epstein et al., 1996; Belhumeur and Kriegman, 1996; Woodham et al., 1991). These approaches rely on studying a single 3D shape cue under the assumptions that other sources of variability can be safely ignored, and that the input photographs contain features relevant to that cue (Bolles and Cain, 1982).[1] Unfortunately, these approaches cannot be easily generalized to attack the N-view reconstruction problem for arbitrary 3D scenes because neither assumption holds true in general. Implicit in this previous work is the view that untangling parallax, self-occlusion and shading effects in N arbitrary photographs of a scene leads to a problem that is either under-constrained or intractable. Here we challenge

this view by showing that shape recovery from N arbitrary photographs of an unknown scene is not only a tractable problem but has a simple solution as well.

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. The Space Carving Algorithm that results from our analysis, however, is related to other 3D scene-space stereo algorithms that have been recently proposed (Fua and Leclerc, 1995; Collins, 1996; Seitz and Dyer, 1999; Seitz and Kutulakos, 1998; Zitnick and Webb, 1996; Narayanan et al., 1998; Szeliski and Golland, 1998; Roy and Cox, 1998). Of these, most closely related are mesh-based (Fua and Leclerc, 1995) and level-set (Faugeras and Keriven, 1998) algorithms, as well as methods that sweep a plane or other manifold through a discretized scene space (Collins, 1996; Seitz and Dyer, 1999; Seitz and Kutulakos, 1998; Szeliski and Golland, 1998; Langer and Zucker, 1994). While the algorithms in (Faugeras and Keriven, 1998; Fua and Leclerc, 1995) 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 Fig. 1(a)). 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 (Collins, 1996; Seitz and Dyer, 1999; Seitz and Kutulakos, 1998; Szeliski and Golland, 1998) are not fully general i.e., they rely on the presence of specific image features such as edges and hence generate only sparse reconstructions (Collins, 1996), or they place strong constraints on the input viewpoints relative to the scene (Seitz and Dyer,1999; Seitz and Kutulakos, 1998). Unlike all previous methods, Space Carving guarantees complete reconstruction in the general case.

Our approach offers six main contributions over the existing state of the art:

1. It introduces an algorithm-independent analysis of the shape recovery problem from N arbitrary photographs, making explicit the assumptions required for solving it as well as the ambiguities intrinsic to the problem. This analysis not only extends previous work on reconstruction but also puts forth a concise geometrical framework for analyzing the general properties of recently-proposed scene-space stereo techniques (Fua and Leclerc, 1995; Collins, 1996; Seitz and Dyer, 1999; Seitz and Kutulakos, 1998; Zitnick and Webb, 1996; Narayanan et al., 1998; Szeliski and Golland, 1998; Roy and Cox, 1998). In this respect, our analysis has goals similar to those of theoretical approaches to structure-from-motion (Faugeras and Maybank, 1990), although the different assumptions employed (i.e., unknown vs. known correspondences, known vs. unknown camera motion), make the geometry, solution space, and underlying techniques completely different.
2. Our analysis provides a volume which is the tightest possible bound on the shape of the true scene that can be inferred from N photographs. This bound is important because it tells us precisely what shape information we can hope to extract from N photographs, in the absence of a priori geometric and point correspondence information, *regardless of the specific algorithm being employed*.
3. The Space Carving Algorithm presented in this paper is the only provably-correct method, to our knowledge, that enables scene reconstruction from input cameras at arbitrary positions. As such, the algorithm enables reconstruction of complex scenes from viewpoints distributed throughout an unknown 3D environment—an extreme example is shown in Fig. 11(a) where the interior and exterior of a house are reconstructed simultaneously from cameras distributed throughout the inside and outside of the house.
4. Because no constraints on the camera viewpoints are imposed, our approach leads naturally to global reconstruction algorithms (Kutulakos and Dyer, 1995; Seitz and Dyer, 1995) that recover 3D shape information from all photographs in a single step. This eliminates the need for complex partial reconstruction and merging operations (Curless and Levoy, 1996; Turk and Levoy, 1994) in which partial 3D shape information is extracted from subsets of the photographs (Narayanan et al., 1998; Kanade et al., 1995; Zhao and Mohr, 1996; Seales and Faugeras, 1995), and where global consistency with the entire set of photographs is not guaranteed for the final shape.
5. We describe an efficient multi-sweep implementation of the Space Carving Algorithm that enables recovery of photo-realistic 3D models from multiple photographs of real scenes, and exploits graphics

hardware acceleration commonly available on desktop PC's.

6. Because the shape recovered via Space Carving is guaranteed to be photo-consistent, its reprojections will closely resemble photographs of the true scene. This property is especially significant in computer graphics, virtual reality, and tele-presence applications (Tomasi and Kanade, 1992; Kanade et al., 1995; Moezzi et al., 1996; Zhang, 1998; Kang and Szeliski, 1996; Sato et al., 1997) where the photo-realism of constructed 3D models is of primary importance.

1.1. Least-Commitment Shape Recovery

A key consequence of our photo-consistency analysis is that there are 3D scenes for which no finite set of input photographs can uniquely determine their shape: in general, there exists an uncountably-infinite equivalence class of shapes each of which reproduces all of the input photographs exactly. This result is yet another manifestation of the well-known fact that 3D shape recovery from a set of images is generally ill-posed (Poggio et al., 1985), i.e., there may be multiple shapes that are consistent with the same set of images.[2] Reconstruction methods must therefore choose a particular scene to reconstruct from the space of all consistent shapes. Traditionally, the most common way of dealing with this ambiguity has been to apply smoothness heuristics and regularization techniques (Poggio et al., 1985; Aloimonos, 1988) to obtain reconstructions that are as smooth as possible. A drawback of this type of approach is that it typically penalizes discontinuities and sharp edges, features that are very common in real scenes.

The notion of the photo hull introduced in this paper and the Space Carving Algorithm that computes it lead to an alternative, *least commitment principle* (Marr, 1982) for choosing among all of the photo-consistent shapes: rather than making an arbitrary choice, we choose the only photo-consistent reconstruction that is guaranteed to subsume (i.e., contain within its volume) all other photo-consistent reconstructions of the scene. By doing so we not only avoid the need to impose ad hoc smoothness constraints, which lead to reconstructions whose relationship to the true shape are difficult to quantify, we also ensure that the recovered 3D shape can serve as a description for the entire equivalence class of photo-consistent shapes.

While our work shows how to obtain a consistent scene reconstruction without imposing smoothness constraints or other geometric heuristics, there are many cases where it may be advantageous to impose a priori constraints, especially when the scene is known to have a certain structure (Debevec et al., 1996; Kakadiaris and Metaxas, 1995). Least-commitment reconstruction suggests a new way of incorporating such constraints: rather than imposing them as early as possible in the reconstruction process, we can impose them after first recovering the photo hull. This allows us to delay the application of a priori constraints until a later stage in the reconstruction process, when tight bounds on scene structure are available and where these constraints are used only to choose among shapes within the class of photo-consistent reconstructions. This approach is similar in spirit to "stratification" approaches of shape recovery (Faugeras, 1995; Koenderink and van Doorn, 1991), where 3D shape is first recovered *modulo* an equivalence class of reconstructions and is then refined within that class at subsequent stages of processing.

The remainder of this paper is structured as follows. Section 2 analyzes the constraints that a set of photographs place on scene structure given a known, locally-computable model of scene radiance. Using these constraints, a theory of photo-consistency is developed that provides a basis for characterizing the space of all reconstructions of a scene. Sections 3 and 4 then use this theory to present the two central results of the paper, namely the existence of the photo hull and the development of a provably-correct algorithm called Space Carving that computes it. Section 5 then presents a discrete implementation of the Space Carving Algorithm that iteratively "carves" out the scene from an initial set of voxels. This algorithm can be seen as a generalization of silhouette-based techniques like volume intersection (Martin and Aggarwal, 1983; Szeliski, 1993; Kutulakos, 1997; Moezzi et al., 1996) to the case of gray-scale and full-color images, and generalizes voxel coloring (Seitz and Dyer, 1999) and plenoptic decomposition (Seitz and Kutulakos, 1998) to the case of arbitrary camera geometries.[3] Section 6 concludes with experimental results on real and synthetic images.

2. Picture Constraints

Let $\mathcal{V}$ be a shape defined by a closed and opaque set of points that occupy a volume in space.[4] We assume that $\mathcal{V}$ is viewed under perspective projection from N known positions $c_1, \ldots, c_N$ in $\mathbb{R}^3 - \mathcal{V}$ (Fig. 1(b)). The *radiance* of a point p on the shape's surface, $Surf(\mathcal{V})$

is a function $rad_p(\xi)$ that maps every oriented ray ξ through the point to the color of light reflected from p along ξ. We use the term *shape-radiance scene description* to denote the shape $\mathcal{V}$ 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.[5]

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*:[6]

Definition 1 (Point Photo-Consistency). Let $\mathcal{S}$ be an arbitrary subset of $\mathbb{R}^3$. A point $p \in \mathcal{S}$ 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 $rad_p(\vec{pc})$. If p is not visible from c, it is trivially photo-consistent with the photograph at c.

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.

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 (Laurentini, 1994).

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.[7] 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 (Szeliski, 1993; Kutulakos, 1997), 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 $\mathbb{R}^3$) or when background identification (Smith and Blinn, 1996) 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 criteria:

Consistency Check Criteria:

1. 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 ξ_i simultaneously for all $i = 1, \ldots, K$.
2. $\texttt{consist}_K()$ is assumed to be monotonic, i.e., $\texttt{consist}_K(col_1, \ldots, col_j, \xi_1, \ldots, \xi_j)$ implies that $\texttt{consist}_K(col_1, \ldots, col_m, \xi_1, \ldots, \xi_m)$ for every $m < j$ and permutation of $1, \ldots, j$.

Given a shape $\mathcal{V}$, the Consistency Check Criteria give us a way to establish the photo-consistency of every point on $\mathcal{V}$'s surface. These criteria define 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 (Szeliski and Golland, 1998), inter-reflection (Forsyth and Zisserman, 1991), and shadows can be ignored. For example, inter-reflection and shadows in Lambertian scenes viewed under fixed illumination are correctly accounted for because scene radiance is isotropic even when such effects are present. As a result, the class of locally-computable radiance models subsumes the Lambertian ($K = 2$) and other parameterized models of scene radiance.[8]

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 (Fig. 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,* $\text{Vis}_{\mathcal{V}}(p) \subseteq \text{Vis}_{\mathcal{V}'}(p)$.

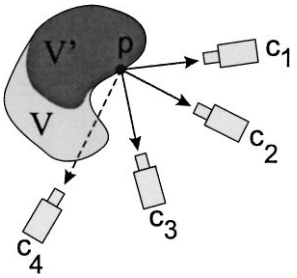


Figure 2. Illustration of the Visibility and Non-Photo-Consistency Lemmas. If p is non-photo-consistent with the photographs at c_1, c_2, c_3, it is non-photo-consistent with the entire set $Vis_{\mathcal{V}'}(p)$, which also includes c_4.

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

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* $\text{Vis}_{\mathcal{V}}(p)$.

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 $p \in Surf(\mathcal{V})$ is visible strictly expands as $\mathcal{V}$ gets smaller (Fig. 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 $p \in$* $\text{Surf}(\mathcal{V})$ *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 $Vis_{\mathcal{V}}(p) \subseteq Vis_{\mathcal{V}'}(p)$. The theorem now follows by applying the Non-Photo-Consistency Lemma to $\mathcal{V}'$ and using the locality property of locally computable radiance models. □

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 (Fig. 3(a)). In general, trivial shape solutions such as this one can 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 set $\mathcal{V}$ within which a photo-consistent shape is known to lie.[9]

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 $\mathcal{V}$ (Fig. 3(b)):

Theorem 2 (*Photo Hull Theorem*)**.** *Let $\mathcal{V}$ be an arbitrary subset of $\mathbb{R}^3$. If $\mathcal{V}^*$ is the union of all photo-consistent shapes in $\mathcal{V}$, every point on the surface of $\mathcal{V}^*$ is photo-consistent. We call $\mathcal{V}^*$ the* photo hull.[10]

Proof: (By contradiction) Suppose that p is a surface point on $\mathcal{V}^*$ that is not photo-consistent. 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. □

Corollary 1. *If $\mathcal{V}^*$ is closed, it is a photo-consistent shape.*

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.

While every point on the photo hull is photo-consistent, the hull itself is not guaranteed to be closed, i.e., it may not satisfy our definition of a *shape*. Specific cases of interest where $\mathcal{V}^*$ is closed include (1) discretized scene volumes, i.e., scenes that are composed of a finite number of volume elements, and (2) instances where the number of photo-consistent shapes in a volume is finite. We describe a volumetric algorithm for computing discretized photo hulls in the next section. The general case, where the photo hull is an infinite union of shapes, is considered in the Appendix.

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 $\mathcal{V}$ 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 (Fig. 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.[11] 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. 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:[12]

Space Carving Algorithm

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

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

- **a.** Project v to all photographs in $Vis_{\mathcal{V}}(v)$. 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:

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.*

Proof: Since (1) the photo-consistency of a voxel v that remains on $\mathcal{V}$'s surface for several carving iterations can change only when $Vis_{\mathcal{V}}(v)$ changes due to $\mathcal{V}$'s carving, and (2) $Vis_{\mathcal{V}}(v)$ expands monotonically as $\mathcal{V}$ is carved (Visibility Lemma), the photo-consistency of v must be checked at most N times. □

5. A Multi-Sweep Implementation of Space Carving

Despite being relatively simple to describe, the Space Carving Algorithm as described in Section 4 requires a difficult update procedure because of the need to keep track of scene visibility from all of the input cameras. In particular, every time a voxel is carved a new set of voxels becomes newly visible and must be re-evaluated for photo-consistency. Keeping track of such changes necessitates computationally-expensive ray-tracing techniques or memory-intensive spatial data structures (Culbertson et al., 1999). To overcome these problems, we instead describe a multi-sweep implementation of the Space Carving Algorithm that enables efficient visibility computations with minimal memory requirements.

5.1. Multi-View Visibility Ordering

A convenient method of keeping track of voxel visibility is to evaluate voxels *in order of visibility*, i.e., visit occluders before the voxels that they occlude. The key advantage of this approach is that backtracking is avoided—carving a voxel affects only voxels encountered later in the sequence. For a single camera, visibility ordering amounts to visiting voxels in a front-to-back order and may be accomplished by depth-sorting (Newell et al., 1972; Fuchs et al., 1980). The problem of defining visibility orders that apply simultaneously to *multiple* cameras is more difficult, however, because it requires that voxels occlude each other in the same order from different viewpoints. More precisely, voxel p is evaluated before q only if q does not occlude p from *any one* of the input viewpoints.

It is known that multi-view visibility orders exist for cameras that lie on one side of a plane (Langer and Zucker, 1994). Recently, Seitz and Dyer (Seitz and Dyer, 1999) generalized this case to a range of interesting camera configurations by showing that multi-view visibility orders always exist when the scene lies outside the convex hull of the camera centers. When this constraint is satisfied, evaluating voxels in order of increasing distance to this camera hull yields a multi-view visibility order that may be used to reconstruct the scene. The convex hull constraint is a significant limitation, however, because it strongly restricts the types of scenes and range of views that are reconstructible. In fact, it can be readily shown that no multi-view visibility constraint exists in general (Fig. 4). Therefore, different techniques are needed in order to reconstruct scenes like Fig. 4 that violate the convex hull constraint.

5.2. Plane-Sweep Visibility

While multi-view visibility orders do not exist in the general case, it *is* possible to define visibility orders that apply to a subset of the input cameras. In particular, consider visiting voxels in order of increasing X coordinate and, for each voxel $p = (X_p, Y_p, Z_p)$, consider only cameras whose X coordinates are less than X_p. If p occludes q from a camera at c, it follows that p is on the line segment cq and therefore $X_p < X_q$. Consequently, p is evaluated before q, i.e., occluders are visited before the voxels that they occlude.

Given this ordering strategy, the Space Carving Algorithm can be implemented as a multi-sweep volumetric algorithm in which a solid block of voxels is iteratively carved away by sweeping a single plane through the scene along a set of pre-defined sweep directions (Fig. 5). For each position of the plane, voxels on the plane are evaluated by considering their projections into input images from viewpoints on one side of the plane. In the example shown in Fig. 5, a plane parallel to the Y-Z axis is swept in the increasing X direction.

Plane Sweep Algorithm

Step 1: Given an initial volume $\mathcal{V}$, initialize the sweep plane Π such that $\mathcal{V}$ lies below Π (i.e., Π is swept towards $\mathcal{V}$).
Step 2: Intersect Π with the current shape $\mathcal{V}$.
Step 3: For each surface voxel v on Π:

- **a.** let $c_1, \ldots, c_j$ be the cameras above Π for which v projects to an *unmarked* pixel;
- **b.** determine the photo-consistency of v using $\texttt{consist}_K(col_1, \ldots, col_j, \xi_1, \ldots, \xi_j)$;
- **c.** if v is inconsistent then set $\mathcal{V} = \mathcal{V} - \{v\}$, otherwise mark the pixels to which v projects.

Step 4: Move Π downward one voxel width and repeat Step 2 until $\mathcal{V}$ lies above Π.

The dominant costs of this algorithm are (1) projecting a plane of voxels into N images, and (2) correlating pixels using $\texttt{consist}_K(col_1, \ldots, col_j, \xi_1, \ldots, \xi_j)$. Our implementation exploits texture-mapping graphics hardware (the kind found on standard PC graphics cards) to project an entire plane of voxels at a time onto each image. We have found that when this optimization is used, the pixel correlation step dominates the computation.

5.3. Multi-Sweep Space Carving

The Plane Sweep Algorithm considers only a subset of the input cameras for each voxel, i.e., the cameras on one side of the sweep plane. Consequently, it may fail to carve voxels that are inconsistent with the entire set of input images but are consistent with a proper subset of these images. To ensure that all cameras are considered, we repeatedly perform six sweeps through the volume, corresponding to the six principle directions (increasing and decreasing X, Y, and Z directions). Furthermore, to guarantee that all cameras visible to a voxel are taken into account, we perform an additional round of voxel consistency checks that incorporate the voxel visibility information collected from individual sweeps. The complete algorithm is as follows:

Multi-Sweep Space Carving Algorithm

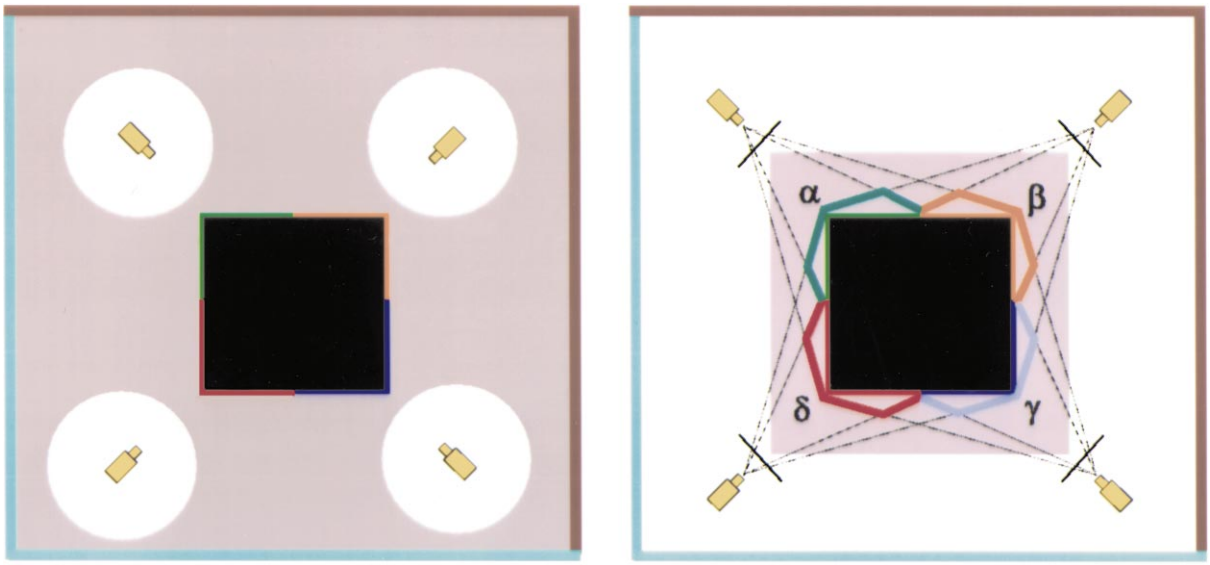
Step 1: Initialize $\mathcal{V}$ to be a superset of the true scene.
Step 2: Apply the Plane Sweep Algorithm in each of the six principle directions and update $\mathcal{V}$ accordingly.
Step 3: For every voxel in $\mathcal{V}$ whose consistency was evaluated in more than one plane sweep:

- **a.** let $c_1, \ldots, c_j$ be the cameras that participated in the consistency check of v in *some* plane sweep during Step 2;
- **b.** determine the photo-consistency of v using $\texttt{consist}_K(col_1, \ldots, col_j, \xi_1, \ldots, \xi_j)$;
- **c.** if v is inconsistent then set $\mathcal{V} = \mathcal{V} - \{v\}$.

Step 4: If no voxels were removed from $\mathcal{V}$ in Steps 2 and 3, set $\mathcal{V}^* = \mathcal{V}$ and terminate; otherwise, repeat Step 2.

5.4. Lambertian Scenes

We give special attention to case of Lambertian scenes, in which the Consistency Check Criteria can be defined using the standard deviation of colors, $col_1, \ldots, col_K$, at a voxel's projection. To account for errors in the image formation process due to quantization, calibration, or other effects, we call a voxel *photo-consistent* if σ is below a given threshold. This threshold is chosen



(a) (b)

Figure 3. Photo-consistent shapes for a two-dimensional scene viewed by four cameras. The scene consists of a black square whose sides are painted diffuse red, blue, orange, and green. (a) Trivial shape solutions in the absence of free-space constraints. Carving out a small circle around each camera and projecting the image onto the interior of that circle yields a trivial photo-consistent shape, shown in gray. (b) Illustration of the Photo Hull Theorem. The gray-shaded region corresponds to an arbitrary shape $\mathcal{V}$ containing the square in (a). $\mathcal{V}^*$ is a polygonal region that extends beyond the true scene and whose boundary is defined by the polygonal segments α, β, γ, and δ. 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 $\mathcal{V}^*$.

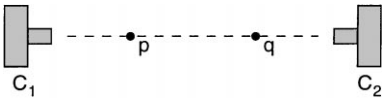
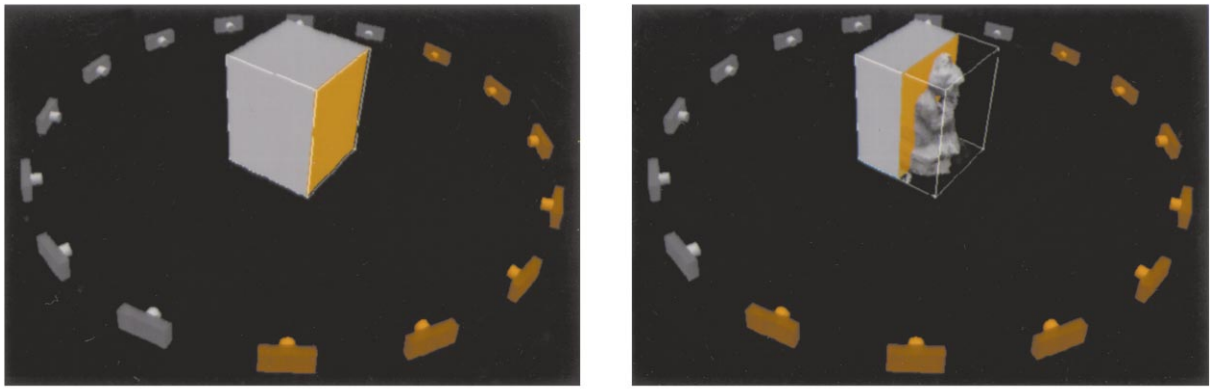


Figure 4. A Visibility Cycle. Voxel p occludes q from c_1, whereas q occludes p from c_2. Hence, no visibility order exists that is the same for both cameras.

by considering σ to be a statistical measure of voxel photo-consistency. In particular, suppose the sensor error (accuracy of irradiance measurement) is normally distributed[13] with standard deviation σ_0. The photo-consistency of a voxel v can be estimated using the likelihood ratio test, distributed as χ^2 with $K-1$ degrees



(a) (b)

Figure 5. Plane-Sweep Visibility. (a) The plane-sweep algorithm ensures that voxels are visited in order of visibility with respect to all active cameras. The current plane and active set of cameras is shown in orange. (b) The shape evolves and new cameras become active as the plane moves through the scene volume.

of freedom (Freund, 1992):

$$\lambda_v = \frac{(K-1)\sigma^2}{\sigma_0^2}. \quad (1)$$

This formulation of the Consistency Check Criterion allows us to incorporate two additional optimizations to the Multi-Sweep Carving Algorithm. First, we maintain sufficient per-voxel color statistics between sweeps to integrate information from all input images, therefore eliminating the need for Step 3 of the multi-sweep algorithm. This is because the standard deviation of K monochrome pixel values of intensity col_i, can be computed using the following recursive formula:

$$\sigma^2 = \frac{1}{K-1}\left(\sum_{i=1}^{K} col_i^2 - \frac{1}{K}\left(\sum_{i=1}^{K} col_i\right)^2\right). \quad (2)$$

It is therefore sufficient to maintain three numbers per voxel, namely $\sum_{i=1}^{K} col_i$, $\sum_{i=1}^{K} col_i^2$, and K (i.e., seven numbers for three-component color pixels). Second, to ensure that no camera is considered more than once per voxel in the six sweeps, we further restrict the cameras considered in each sweep to a pyramidal beam defined by the voxel center and one of its faces, as shown in Fig. 6. This strategy partitions the cameras into six non-overlapping sets to be processed in the six respective sweeps, thereby ensuring that each camera is considered exactly once per voxel during the six sweeps.

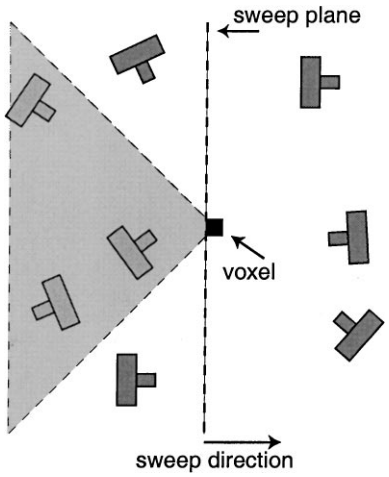


Figure 6. To ensure that a camera is processed at most once per voxel during the six plane sweeps, the set of cameras considered in each sweep is clipped to a pyramidal beam defined by the center of the voxel and one of its faces.

6. 3D Photography by Space Carving

6.1. Image Acquisition

In the Space Carving Algorithm, every input photograph can be thought of as a *shape constraint* that forces the reconstructed scene volume to contain only voxels consistent with the photograph. To ensure that the algorithm's output closely resembles the shape and appearance of a complicated 3D scene it is therefore important to acquire enough photographs of the scene itself. In a typical image acquisition session, we take between 10 and 100 calibrated images around the scene of interest using a Pulnix TMC-9700 color CCD camera (Fig. 7).

A unique property of the Space Carving Algorithm is that it can be forced to automatically segment a 3D object of interest from a larger scene using two complementary methods. The first method, illustrated in the sequence of Fig. 7, involves slightly modifying the image acquisition process—before we take a photograph of the object of interest from a new viewpoint, we manually alter the object's background. This process enabled segmentation and complete reconstruction of the gargoyle sculpture; the Space Carving Algorithm effectively removed all background pixels in all input photographs because the varying backgrounds ensured that photo-consistency could not be enforced for points projecting to non-object pixels. Note that image subtraction or traditional matting techniques (Smith and Blinn, 1996) cannot be applied to this image sequence to segment the sculpture since every photograph was taken from a *different* position in space and therefore the background is different in each image. The second method, illustrated in Fig. 9, involves defining an initial volume $\mathcal{V}$ (e.g., a bounding box) that is "tight enough" to ensure reconstruction of only the object of interest. This process enabled segmentation of the hand because the initial volume did not intersect distant objects such as the TV monitor.

6.2. Reconstruction Results

In this section we present results from applying our Multi-Sweep implementation of the Space Carving Algorithm to a variety of 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 the same color in every image. The

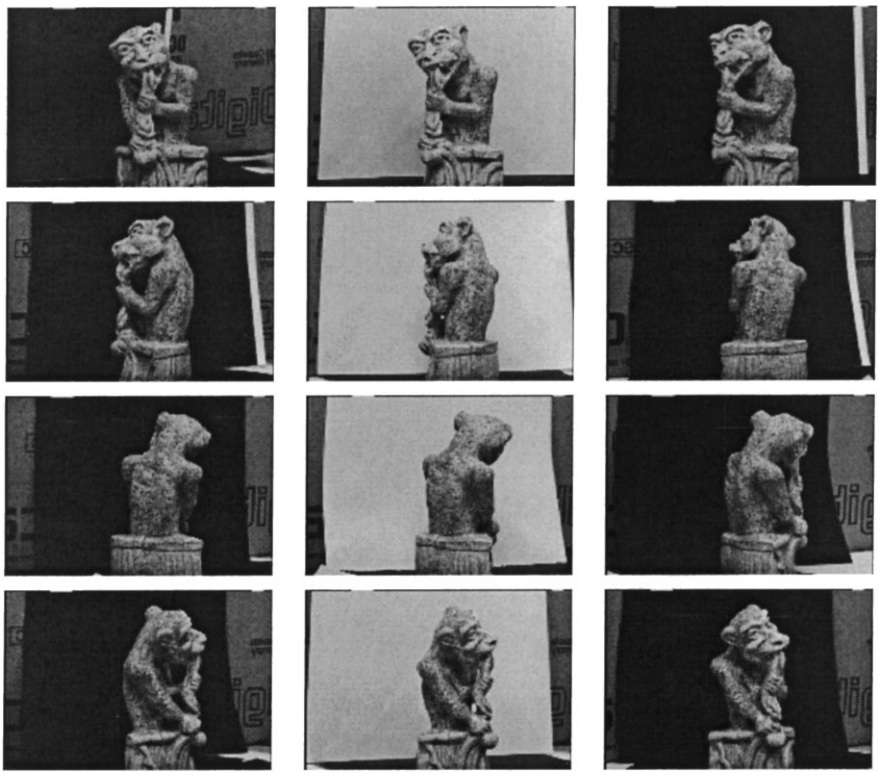


Figure 7. Nine of sixteen 486 × 720 RGB images of a gargoyle stone sculpture. The sequence corresponds to a complete circumnavigation of the object, performed in approximately 22.5 degree increments.

standard deviation of these pixels was therefore used to determine whether or not a voxel should be carved, as described in Section 5.

We first ran the Space Carving Algorithm on 16 images of a gargoyle sculpture (Fig. 7). 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. Figure 8 shows selected input images and new views of the reconstruction. This reconstruction consisted of 215 thousand surface voxels that were carved out of an initial volume of approximately 51 million voxels. It took 250 minutes to compute on an SGI O2 R10000/175 MHz workstation. Some errors are still present in the reconstruction, notably holes that occur as a result of shadows and other illumination changes due to the object's rotation inside a static, mostly diffuse illumination environment. These effects were not modeled by the Lambertian model and therefore caused voxels on shadowed surfaces to be 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.

Results from a sequence of one hundred images of a hand are shown in Figs. 9 and 10. Note that the near-perfect segmentation of the hand from the rest of the scene was performed not in image-space, but in 3D object space—the background lay outside the initial block of voxels and was therefore not reconstructed. This method of 3D background segmentation has significant advantages over image subtraction and chroma-keying methods because it (1) does not require the background to be known and (2) will never falsely eliminate foreground pixels, as these former techniques are prone to do (Smith and Blinn, 1996).

Two kinds of artifacts exist in the resulting reconstructions. First, voxels that are not visible from any input viewpoint do not have a well-defined color assignment and are given a default color. These artifacts

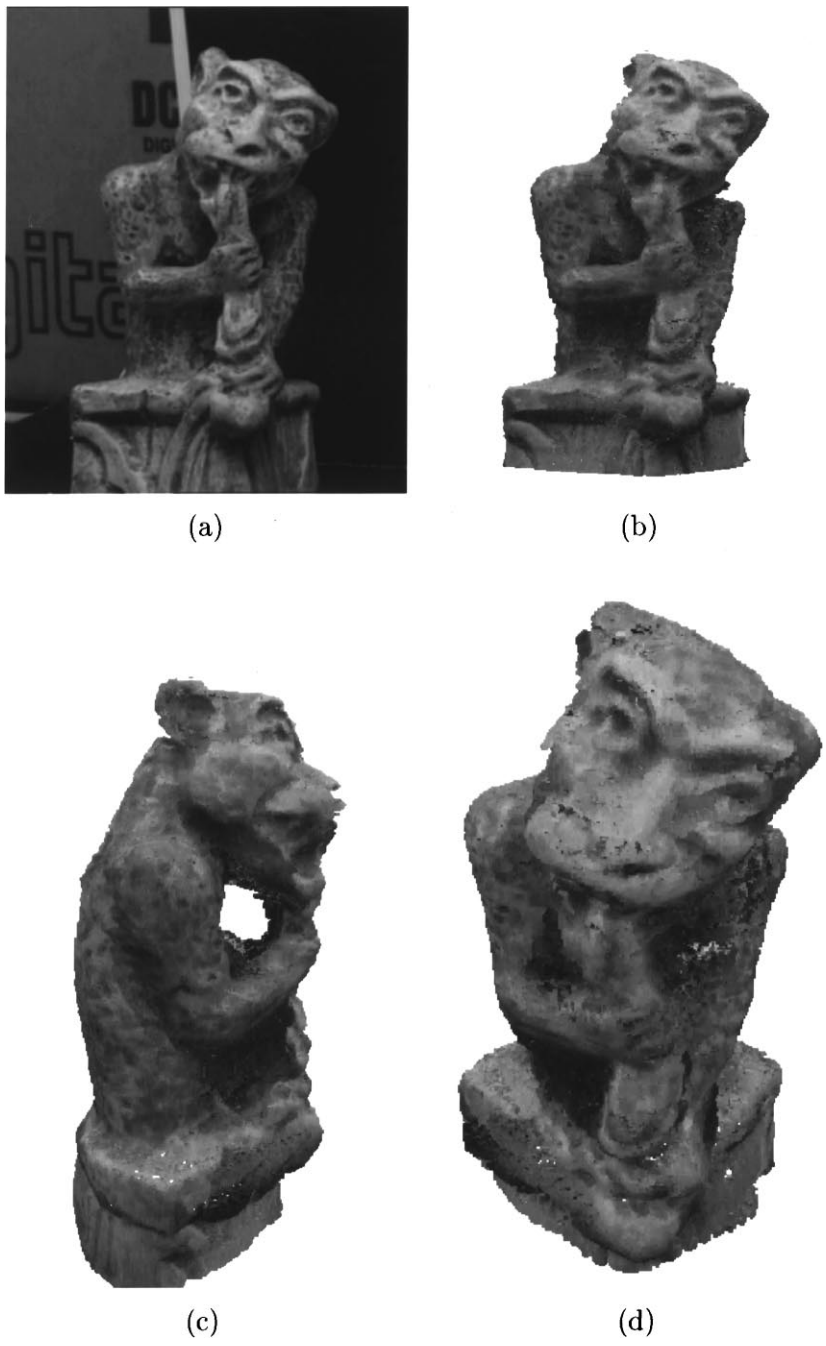


Figure 8. Reconstruction of a gargoyle sculpture. One of 16 input images is shown (a), along with views of the reconstruction from the same (b) and new (c)–(d) viewpoints.

can be eliminated by acquiring additional photographs to provide adequate coverage of the scene's surfaces. Second, stray voxels may be reconstructed in unoccupied regions of space due to accidental agreements between the input images. Such artifacts can be easily avoided by re-applying the Space Carving Algorithm on an initial volume that does not contain those regions or by post-filtering the reconstructed voxel model.

In a final experiment, we applied our algorithm to images of a synthetic building scene rendered from

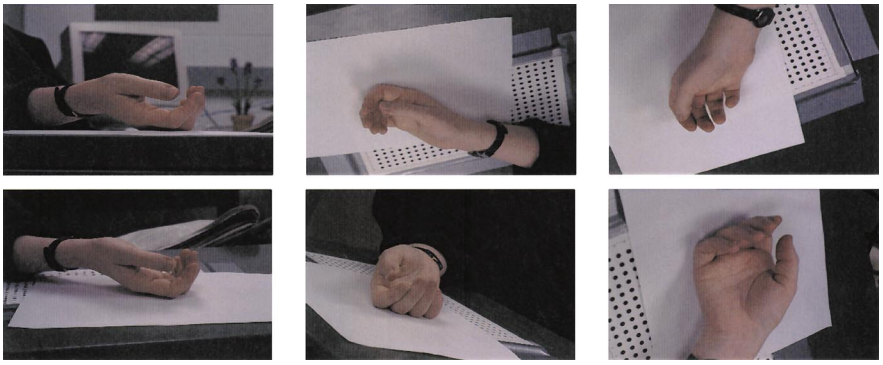


Figure 9. Six out of one hundred photographs of a hand sequence.

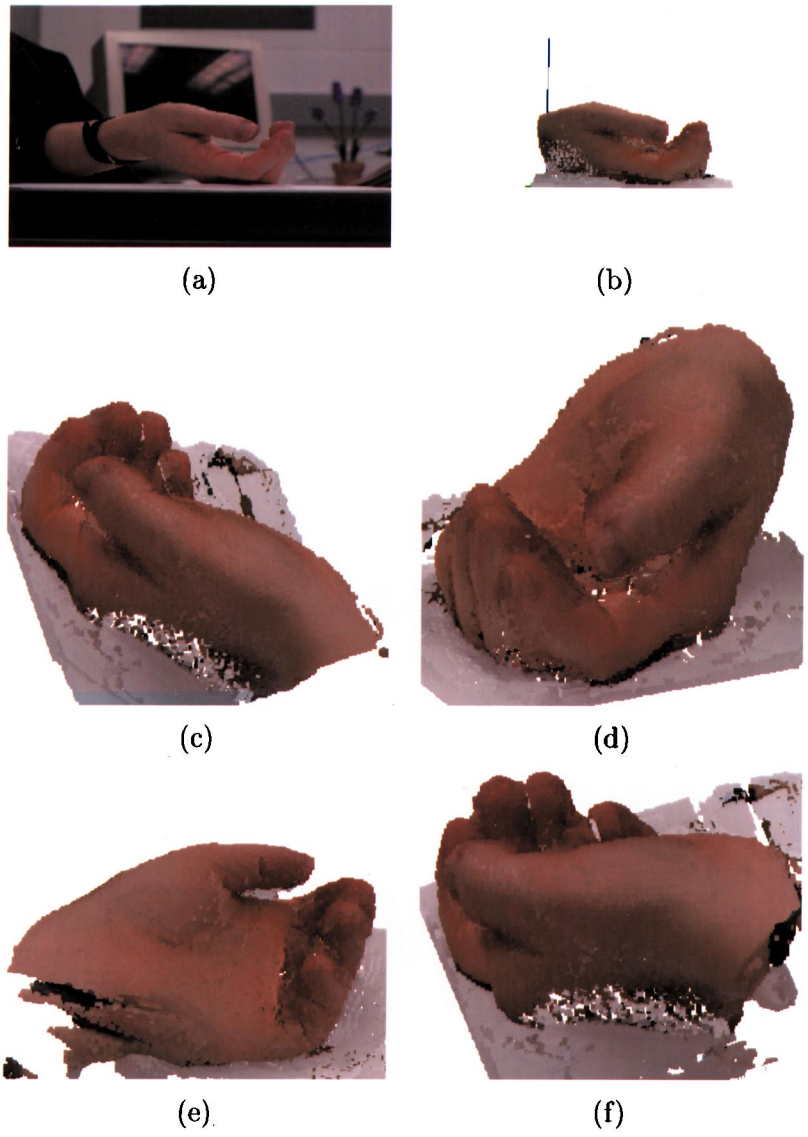


Figure 10. Reconstruction of a hand. An input image is shown in (a) along with views of the reconstruction from the same (b) and other (c)–(f) viewpoints. The reconstructed model was computed using an RGB component error threshold of 6%. The model has 112 thousand voxels and took 53 minutes to compute. The blue line in (b) indicates the z-axis of the world coordinate system.

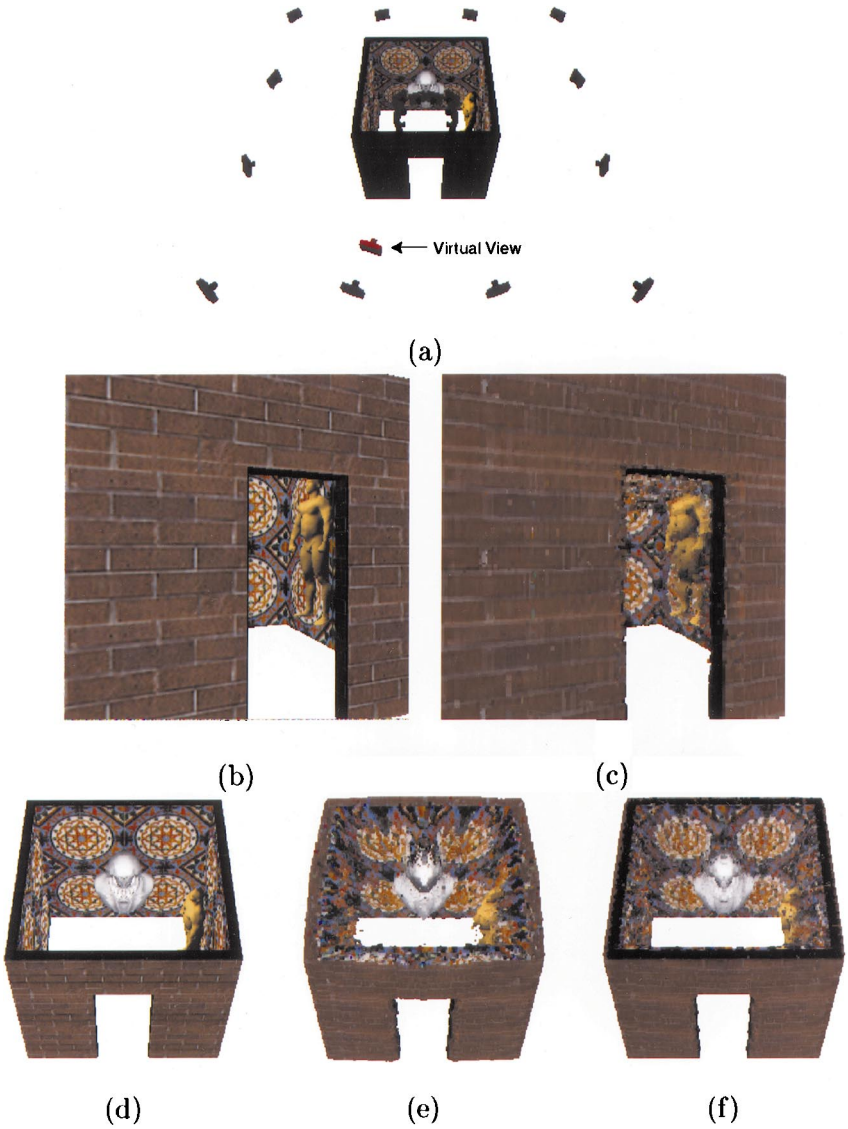


Figure 11. Reconstruction of a synthetic building scene. (a) 24 Cameras were placed in both the interior and exterior of a building to enable simultaneous, complete reconstruction of its exterior and interior surfaces. The reconstruction contains 370,000 voxels, carved out of a $200 \times 170 \times 200$ 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.

both its interior and exterior (Fig. 11). This placement of cameras yields an extremely difficult stereo problem, due to the drastic changes in visibility between interior and exterior cameras.[14] Figure 11 compares the original model and the reconstruction from different viewpoints. The model's appearance is very good near the input viewpoints, as demonstrated in Fig. 11(b)–(c). Note that the reconstruction tends to "bulge" out and that the walls are not perfectly planar (Fig. 11(e)). 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 Fig. 11(f).

Our experiments highlight a number of advantages of our approach over previous techniques. Existing multi-baseline stereo techniques (Okutomi and Kanade, 1993) work best for densely textured scenes and suffer in the presence of large occlusions. In contrast, the hand sequence contains many low-textured regions and dramatic changes in visibility. The low-texture and occlusion properties of such scenes cause problems for feature-based structure-from-motion methods (Tomasi and Kanade, 1992; Seitz and Dyer, 1995; Beardsley et al., 1996; Pollefeys et al., 1998), due to the difficulty of locating and tracking a sufficient number of features throughout the sequence. While contour-based techniques like volume intersection (Martin and Aggarwal, 1983; Szeliski, 1993) often 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 (Seitz and Dyer, 1999) 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.

7. Concluding Remarks

This paper introduced *photo-consistency theory* as a new, general mathematical framework for analyzing the 3D shape recovery problem from multiple images. We have shown that this theory leads to a "least commitment" approach for shape recovery and a practical algorithm called Space Carving that together overcome several limitations in the current state of the art. First, the approach allows us to analyze and characterize the set of all possible reconstructions of a scene, without placing constraints on geometry, topology, or camera configuration. Second, this is the only provably-correct method, to our knowledge, capable of reconstructing non-smooth, free-form shapes from cameras positioned and oriented in a completely arbitrary way. Third, the performance of the Space Carving Algorithm was demonstrated on real and synthetic image sequences of geometrically-complex objects, including a large building scene photographed from both interior and exterior viewpoints. Fourth, the use of photo-consistency as a criterion for 3D shape recovery enables the development of reconstruction algorithms that allow faithful image reprojections and resolve the complex interactions between occlusion, parallax, and shading effects in shape analysis.

While the Space Carving Algorithm's effectiveness was demonstrated in the presence of low 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 (Kutulakos, 2000). Extending the formulation to handle non-locally computable radiance models (e.g., shadows and inter-reflections) is another important topic of future work. Other research directions include (1) developing space carving algorithms for images with significant pixel noise, (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. Finally, an on-line implementation of the Space Carving Algorithm, that performs image capture and scene reconstruction simultaneously, would be extremely useful both to facilitate the image acquisition process and to eliminate the need to store long video sequences.

Appendix

In general, the photo hull, $\mathcal{V}^*$, of a set $\mathcal{V}$ is the union of a potentially infinite collection of shapes in $\mathcal{V}$. Such a union does not always correspond to a closed subset of $\mathbb{R}^3$ (Armstrong, 1983). As a result, even though all points of the photo hull are photo-consistent, the photo hull itself may not satisfy the definition of a 3D shape given in Section 2. In this Appendix we investigate the properties of the closure, $\overline{\mathcal{V}^*}$, of $\mathcal{V}^*$ which is always a valid shape.[15] In particular, we show that $\overline{\mathcal{V}^*}$ satisfies a slightly weaker form of photo-consistency called *directional ϵ-photo-consistency*, defined below. This property leads to a generalization of Theorem 2:

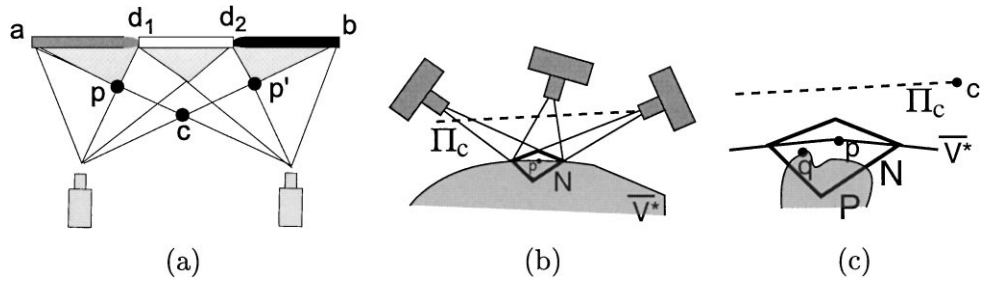


Figure 12. (a) Non-photo-consistent points on the closed photo hull. The 2D scene is composed of a closed thick line segment ab that is painted gray, white, and black. The points d_1, d_2, corresponding to color transitions, are painted white. When $\mathcal{V}$ is defined by the triangle abc, the closed photo hull, $\overline{\mathcal{V}}^*$, is defined by the region shown in light gray. Note that even though $p \in \overline{\mathcal{V}}^*$ is directionally ϵ-photo-consistent, it is not photo-consistent: p projects to a white pixel in the left camera and a gray pixel in the right one. (b)–(c) Proof of Theorem 3. (b) A point p is strongly visible to three cameras by means of neighborhood N. (c) The closest point $q \in N \cap \mathcal{P}$ to Π_c is visible to all cameras on or above Π_c.

Theorem 3 (*Closed Photo Hull Theorem*)**.** *Let $\mathcal{V}$ be an arbitrary shape in $\mathbb{R}^3$ and let $\overline{\mathcal{V}}^*$ be the closure of the union of all photo-consistent shapes in $\mathcal{V}$. The shape $\overline{\mathcal{V}}^*$ is directionally ϵ-photo-consistent and is called the* closed photo hull.

A.1. The Strong Visibility Condition

Because we impose no constraints on the structure of the photo-consistent shapes in $\mathcal{V}$ that are considered in our analysis (e.g., smoothness), it is possible to define degenerate shapes that defy one's "intuitive" notions of visibility and occlusion. More specifically, the standard definition of visibility of a surface point p from a camera c requires that the open line segment pc does not intersect the shape itself; otherwise, p is defined to be occluded. When $\mathcal{V}$ is arbitrary, however, it is possible to define shapes whose surface gets infinitesimally close to this line segment at one or more points other than p. Intuitively, surface points that have this property are not occluded under the above definition but are not "fully visible" either. We therefore refine the notion of visibility in a way that excludes such degeneracies. In particular, let $B(p, \epsilon) \subset \mathbb{R}^3$ be the open 3-ball of radius ϵ that is centered at p:

Definition 4 (Strong Visibility Condition). A point p on the surface of a shape $\mathcal{V}$ is strongly visible to a set of cameras if it is visible from those cameras and if, for every $\epsilon > 0$, there exists a closed set N and an $\epsilon' < \epsilon$ such that the following two properties are satisfied:

1. N contains all its occluders, i.e., for every camera c and point $p \in N$, if q occludes p from c then $q \in N$, and
2. $B(p, \epsilon') \subset N \subset B(p, \epsilon)$.

Intuitively, the strong visibility condition is equivalent to the standard definition of point visibility for shapes that are "well-behaved"—it differs from this definition only in cases where the ray from point p to a camera comes arbitrarily close to the shape outside p's neighborhood. An illustration of a strong visibility neighborhood N is given in Fig. 12(b).

A.2. Directional ϵ-Photo-Consistency

When $\overline{\mathcal{V}}^*$ and $\mathcal{V}^*$ are not equal, the closed photo hull will contain limit points that do not belong to any photo-consistent subset of $\mathcal{V}$. These limit points are not always photo-consistent (Fig. 12(a)). Fortunately, even though the photo-consistency of these points cannot be guaranteed, these points (as well as the rest of $\overline{\mathcal{V}}^*$) do satisfy the directional ϵ-photo-consistency property:

Definition 5 (Strongly Visible Camera Set). If $p \in \mathcal{V}$, Π_p is a plane through p, and C is the set of cameras in $Vis_{\mathcal{V}}(p)$ that are strictly above Π_p, define

$$SVis_{\mathcal{V}}(\Pi_p) = \begin{cases} C & \text{if } p \text{ is strongly visible to } C, \\ \emptyset & \text{otherwise.} \end{cases} \quad (3)$$

Definition 6 (Directional Point Photo-Consistency). A point p in $\mathcal{V}$ is directionally photo-consistent if for every oriented plane Π_p through p, the point p is photo-consistent with all cameras in $SVis_{\mathcal{V}}(\Pi_p)$.

Definition 7 (Directional ϵ-photo-consistency). A point p in $\mathcal{V}$ is directionally ϵ-photo-consistent if for every $\epsilon > 0$ and every oriented plane Π_p through p,

there exists a point $q \in B(p, \epsilon)$ that is photo-consistent with all cameras in $SVis_{\mathcal{V}}(\Pi_p)$.

Compared to the definition of point photo-consistency (Definition 1), directional photo-consistency relaxes the requirement that p's radiance assignment must agree with all visible cameras. Instead, it requires the ability to find radiance assignment(s) that force agreement only with visible cameras within the same half-space. Directional ϵ-photo-consistency goes a step further, lifting the requirement that every surface point p must have a directionally consistent radiance assignment. The only requirement is that p is infinitesimally close to a point for which directional consistency can be established with respect to the cameras from which p is strongly visible.

Despite their differences, photo-consistency and directional ϵ-photo-consistency share a common characteristic: we can determine whether or not these properties hold for a given shape $\mathcal{V}$ without having *any* information about the photo-consistent shapes contained in $\mathcal{V}$. This is especially important when attempting to characterize $\overline{\mathcal{V}}^*$ because it establishes a direct link between $\overline{\mathcal{V}}^*$ and the image observations that does not depend on explicit knowledge of the family of photo-consistent shapes.

A.3. Proof of Theorem 3

Since points that are not strongly visible are always ϵ-photo-consistent, it is sufficient to consider only strongly visible points $p \in \overline{\mathcal{V}}^*$. More specifically, it suffices to show that every open ball, $B(p, \epsilon)$, contains a point q on some photo-consistent shape $\mathcal{P}$ such that the set $Vis_{\mathcal{P}}(q)$ contains all cameras in $SVis_{\overline{\mathcal{V}}^*}(\Pi_p)$. For if q is photo-consistent with $Vis_{\mathcal{P}}(q)$, it follows that q is photo-consistent with any of its subsets.

We proceed by first choosing a photo-consistent shape $\mathcal{P}$ and then constructing the point q (Fig. 12(b) and (c)). In particular, let c be a camera in $SVis_{\overline{\mathcal{V}}^*}(\Pi_p)$ that is closest to Π_p, and let Π_c be the plane through c that is parallel to Π_p. Fix ϵ such that $0 < \epsilon < k$, where k is the distance from c to Π_p.

Let $N \subset B(p, \epsilon)$ be a set that establishes p's strong visibility according to Definition 4. According to the definition, N contains an open ball $B(p, \epsilon')$ for some $\epsilon' < \epsilon$. By the definition of the photo hull, there exists a photo-consistent shape $\mathcal{P}$ that intersects $B(p, \epsilon')$.

We now construct point q and consider the set of cameras from which q is visible. Let q be a point in the set $\mathcal{P} \cap N$ that minimizes perpendicular distance to Π_c.[16] By construction, no point in $N \cap \mathcal{P}$ occludes q from the cameras in $SVis_{\overline{\mathcal{V}}^*}(\Pi_p)$. Moreover, since $q \in N$, Definition 4 tells us that no point in $\mathcal{P} - N$ can occlude q from the cameras in $SVis_{\overline{\mathcal{V}}^*}(\Pi_p)$. It follows that $Vis_{\mathcal{P}}(q) \supseteq SVis_{\overline{\mathcal{V}}^*}(\Pi_p)$. □

Acknowledgments

The authors would like to thank Prof. Terry Boult for many discussions and suggestions on the technical aspects of the paper's proofs. Kiriakos Kutulakos gratefully acknowledges the support of the National Science Foundation under Grant No. IIS-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.

Notes

1. Examples include the use of the small baseline assumption in stereo to simplify correspondence-finding and maximize joint visibility of scene points (Kanade et al., 1996), the availability of easily-detectable image contours in shape-from-contour reconstruction (Vaillant and Faugeras, 1992), and the assumption that all views are taken from the same viewpoint in photometric stereo (Woodham et al., 1991).
2. Faugeras (Faugeras, 1998) has recently proposed the term *metameric* to describe such shapes, in analogy with the term's use in the color perception (Alfvin and Fairchild, 1997) and structure-from-motion literature (van Veen and Werkhoven, 1996).
3. Note that both of these generalizations represent significant improvements in the state of the art. For instance, silhouette-based algorithms require identification of silhouettes, fail at surface concavities, and treat only the case of binary images. While (Seitz and Dyer, 1999; Seitz and Kutulakos, 1998) also used a volumetric algorithm, their method worked only when the scene was outside the convex hull of the cameras. This restriction strongly limits the kinds of environments that can be reconstructed, as discussed in Section 6.
4. More formally, we use the term *shape* to refer to any closed set $\mathcal{V} \subseteq \mathbb{R}^3$ for which every point $p \in \mathcal{V}$ is infinitesimally close to an open 3-ball inside $\mathcal{V}$. That is, for every $\epsilon > 0$ there is an open 3-ball, $B(p, \epsilon)$, that contains an open 3-ball lying inside $\mathcal{V}$. Similarly, we define the *surface* of $\mathcal{V}$ to be the set of points in $\mathcal{V}$ that are infinitesimally close to a point outside $\mathcal{V}$.
5. Note that even points on a radiance discontinuity must have a unique radiance function assigned to them. For example, in the scene of Fig. 3, the point of transition between red and blue surface points must be assigned either a red or a blue color.
6. In the following, we make the simplifying assumption that pixel values in the image measure scene radiance directly.

7. For example, set $rad_p(\vec{pc})$ equal to the color at p's projection.
8. Strictly speaking, locally-computable radiance models cannot completely account for surface normals and other neighborhood-dependent quantities. However, it is possible to estimate surface normals based purely on radiance information and thereby approximately model cases where the light source changes (Seitz and Kutulakos, 1998) or when reflectance is normal-dependent (Sato et al., 1997). 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 (Torrance and Sparrow, 1967; Sato et al., 1997)), and (2) using multiple cameras to capture photographs of an approximately Lambertian scene under arbitrary unknown illumination (Fig. 1).
9. Note that if $\mathcal{V} = \mathbb{R}^3$, the problem reduces to the case when no constraints on free space are available.
10. 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.
11. Examples include defining $\mathcal{V}$ to be equal to the visual hull or, in the case of a camera moving through an environment, $\mathbb{R}^3$ minus a tube along the camera's path.
12. Convergence to this shape is provably guaranteed only for scenes representable by a discrete set of voxels.
13. Here we make the simplifying assumption that σ_0 does not vary as a function of wavelength.
14. For example, the algorithms in (Seitz and Dyer, 1999; Seitz and Kutulakos, 1998) fail catastrophically for this scene because the distribution of the input views and the resulting occlusion relationships violate the assumptions used by those algorithms.
15. To see this, note that $\overline{\mathcal{V}}^*$ is, by definition, a closed subset of $\mathbb{R}^3$. Now observe that every point $p \in \overline{\mathcal{V}}^*$ is infinitesimally close to a point on *some* photo-consistent shape $\mathcal{V}'$. It follows that p is infinitesimally close to an open 3-ball inside $\mathcal{V}' \subseteq \overline{\mathcal{V}}^*$. The closed photo hull therefore satisfies our definition of a shape.
16. Note that such a point does exist since $\mathcal{P} \cap N$ is a closed and bounded subset of $\mathbb{R}^3$ and hence it is compact (Armstrong, 1983).

References

Alfvin, R.L. and Fairchild, M.D. 1997. Observer variability in metameric color matches using color reproduction media. *Color Research & Application*, 22(3):174–178.

Aloimonos, Y. 1988. Visual shape computation. *Proc. IEEE*, 76:899–916.

Armstrong, M.A. 1983. *Basic Topology*. Springer-Verlag.

Bascle, B. and Deriche, R. 1993. Stereo matching, reconstruction and refinement of 3D curves using deformable contours. In *Proc. 4th Int. Conf. Computer Vision*, pp. 421–430.

Beardsley, P., Torr, P., and Zisserman, A. 1996. 3D model acquisition from extended image sequences. In *Proc. 4th European Conf. on Computer Vision*, pp. 683–695.

Belhumeur, P.N. 1996. A bayesian approach to binocular stereopsis. *Int. J. on Computer Vision*, 19(3):237–260.

Belhumeur, P.N. and Kriegman, D.J. 1996. What is the set of images of an object under all possible lighting conditions? In *Proc. Computer Vision and Pattern Recognition*, pp. 270–277.

Bolles, R.C., Baker, H.H., and Marimont, D.H. 1987. Epipolar-plane image analysis: An approach to determining structure from motion. *Int. J. Computer Vision*, 1:7–55.

Bolles, R.C. and Cain, R.A. 1982. Recognizing and locating partially-visible objects: The local-feature-focus method. *Int. J. Robotics Research*, 1(3):57–82.

Cipolla, R. and Blake, A. 1992. Surface shape from the deformation of apparent contours. *Int. J. Computer Vision*, 9(2):83–112.

Collins, R.T. 1996. A space-sweep approach to true multi-image matching. In *Proc. Computer Vision and Pattern Recognition Conf.*, pp. 358–363.

Cox, I., Hingorani, S., Rao, S., and Maggs, B. 1996. A maximum likelihood stereo algorithm. *CVIU: Image Understanding*, 63(3):542–567.

Culbertson, W.B., Malzbender, T., and Slabaugh, G. 1999. Generalized voxel coloring. In *Workshop on Vision Algorithms: Theory and Practice*, Corfu, Greece.

Curless, B. and Levoy, M. 1996. A volumetric method for building complex models from range images. In *Proc. SIGGRAPH'96*, pp. 303–312.

Debevec, P.E., Taylor, C.J., and Malik, J. 1996. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *Proc. SIGGRAPH'96*, pp. 11–20.

Epstein, R., Yuille, A.L., and Belhumeur, P.N. 1996. Learning object representations from lighting variations. In *Object Representation in Computer Vision II*, J. Ponce, A. Zisserman, and M. Hebert (Eds.). Springer-Verlag, pp. 179–199.

Faugeras, O. 1995. Stratification of three-dimensional vision: Projective, affine, and metric representations. *J. Opt. Soc. Am. A*, 12(3):465–484.

Faugeras, O.D. 1998. Personal communication.

Faugeras, O. and Keriven, R. 1998. Complete dense stereovision using level set methods. In *Proc. 5th European Conf. on Computer Vision*, pp. 379–393.

Faugeras, O.D. and Maybank, S. 1990. Motion from point matches: Multiplicity of solutions. *Int. J. Computer Vision*, 4:225–246.

Foley, J.D., van Dam, A., Feiner, S.K., and Hughes, J.F. 1990. *Computer Graphics Principles and Practice*. Addison-Wesley Publishing Co.

Forsyth, D. and Zisserman, A. 1991. Reflections on shading. *IEEE Trans. Pattern Anal. Machine Intell.*, 13(7):671–679.

Freund, J.E. 1992. *Mathematical Statistics*. Prentice Hall: Englewood Cliffs, NJ.

Fua, P. and Leclerc, Y.G. 1995. Object-centered surface reconstruction: Combining multi-image stereo and shading. *Int. J. Computer Vision*, 16:35–56.

Fuchs, H., Kedem, Z., and Naylor, B.F. 1980. On visible surface generation by a priori tree structures. In *Proc. SIGGRAPH '80*, pp. 39–48.

Hoff, W. and Ahuja, N. 1989. Surfaces from stereo: Integrating feature matching, disparity estimation, and contour detection. *IEEE Trans. Pattern Anal. Machine Intell.*, 11:121–136.

Kakadiaris, I.A. and Metaxas, D. 1995. 3D human body model acquisition from multiple views. In *Proc. Int. Conf. on Computer Vision*, pp. 618–623.

Kanade, T., Narayanan, P.J., and Rander, P.W. 1995. Virtualized reality: Concepts and early results. In *Proc. Workshop on Representations of Visual Scenes*, pp. 69–76.

Kanade, T., Yoshida, A., Oda, K., Kano, H., and Tanaka, M. 1996. A stereo machine for video-rate dense depth mapping and its new

applications. In *Proc. Computer Vision and Pattern Recognition Conf.*

Kang, S.B. and Szeliski, R. 1996. 3-D scene data recovery using omnidirectional multibaseline stereo. In *Proc. Computer Vision and Pattern Recognition Conf.*, pp. 364–370.

Katayama, A., Tanaka, K., Oshino, T., and Tamura, H. 1995. A viewpoint dependent stereoscopic display using interpolation of multi-viewpoint images. In *Proc. SPIE*, Vol. 2409A, pp. 21–30.

Koenderink, J.J. and van Doorn, A.J. 1991. Affine structure from motion. *J. Opt. Soc. Am.*, A(2):377–385.

Kutulakos, K.N. 1997. Shape from the light field boundary. In *Proc. Computer Vision and Pattern Recognition*, pp. 53–59.

Kutulakos, K.N. 2000. Approximate N-view stereo. In *Proc. European Conf. on Computer Vision*.

Kutulakos, K.N. and Dyer, C.R. 1994. Recovering shape by purposive viewpoint adjustment. *Int. J. Computer Vision*, 12(2):113–136.

Kutulakos, K.N. and Dyer, C.R. 1995. Global surface reconstruction by purposive control of observer motion. *Artificial Intelligence Journal*, 78(1–2):147–177.

Langer, M.S. and Zucker, S.W. 1994. Shape-from-shading on a cloudy day. *J. Opt. Soc. Am. A*, 11(2):467–478.

Laurentini, A. 1994. The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Anal. Machine Intell.*, 16(2):150–162.

Marr, D. 1982. *Vision*. Freeman.

Martin, W.N. and Aggarwal, J.K. 1983. Volumetric descriptions of objects from multiple views. *IEEE Proc. Pattern Anal. Machine Intell.*, 5(2):150–158.

Moezzi, S., Katkere, A., Kuramura, D.Y., and Jain, R. 1996. Reality modeling and visualization from multiple video sequences. *IEEE Computer Graphics and Applications*, 16(6):58–63.

Mundy, J.L. and Zisserman, A. (Eds.). 1992. *Geometric Invariance in Computer Vision*. MIT Press.

Narayanan, P.J., Rander, P.W., and Kanade, T. 1998. Constructing virtual worlds using dense stereo. In *Proc. Int. Conf. on Computer Vision*, pp. 3–10.

Newell, M.E., Newell, R.G., and Sancha, T.L. 1972. A solution to the hidden surface problem. In *Proc. ACM National Conference*, pp. 443–450.

Okutomi, M. and Kanade, T. 1993. A multiple-baseline stereo. *IEEE Trans. Pattern Anal. Machine Intell.*, 15(4):353–363.

Poggio, T., Torre, V., and Koch, C. 1985. Computational vision and regularization theory. *Nature*, 317(26):314–319.

Pollefeys, M., Koch, R., and Gool, L.V. 1998. Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. In *Proc. 6th Int. Conf. on Computer Vision*, pp. 90–95.

Pritchett, P. and Zisserman, A. 1998. Wide baseline stereo matching. In *Proc. 6th Int. Conf. on Computer Vision*, pp. 754–760.

Roy, S. and Cox, I.J. 1998. A maximum-flow formulation of the N-camera stereo correspondence problem. In *Proc. 6th Int. Conf. on Computer Vision*, pp. 492–499.

Sato, Y., Wheeler, M.D., and Ikeuchi, K. 1997. Object shape and reflectance modeling from observation. In *Proc. SIGGRAPH'97*, pp. 379–387.

Seales, W.B. and Faugeras, O. 1995. Building three-dimensional object models from image sequences. *Computer Vision and Image Understanding*, 61(3):308–324.

Seitz, S.M. and Dyer, C.R. 1995. Complete scene structure from four point correspondences. In *Proc. 5th Int. Conf. on Computer Vision*, pp. 330–337.

Seitz, S.M. and Dyer, C.R. 1999. Photorealistic scene reconstruction by voxel coloring. *Int. J. Computer Vision*, 35(2):151–173.

Seitz, S.M. and Kutulakos, K.N. 1998. Plenoptic image editing. In *Proc. 6th Int. Conf. Computer Vision*, pp. 17–24.

Smith, A.R. and Blinn, J.F. 1996. Blue screen matting. In *Proc. SIGGRAPH'96*, pp. 259–268.

Stewart, C.V. 1995. MINPRAN: A new robust estimator for computer vision. *IEEE Trans. Pattern Anal. Machine Intell.*, 17(10): 925–938.

Szeliski, R. 1993. Rapid octree construction from image sequences. *CVGIP: Image Understanding*, 58(1):23–32.

Szeliski, R. and Golland, P. 1998. Stereo matching with transparency and matting. In *Proc. 6th Int. Conf. on Computer Vision*, pp. 517–524.

Szeliski, R. and Weiss, R. 1994. Robust shape recovery from occluding contours using a linear smoother. In *Real-time Computer Vision*, C.M. Brown and D. Terzopoulos (Eds.). Cambridge University Press, pp. 141–165.

Tomasi, C. and Kanade, T. 1992. Shape and motion from image streams under orthography: A factorization method. *Int. J. Computer Vision*, 9(2):137–154.

Torrance, K.E. and Sparrow, E.M. 1967. Theory of off-specular reflection from roughened surface. *Journal of the Optical Society of America*, 57:1105–1114.

Turk, G. and Levoy, M. 1994. Zippered polygon meshes from range images. In *Proc. SIGGRAPH'94*, pp. 311–318.

Vaillant, R. and Faugeras, O.D. 1992. Using extremal boundaries for 3-D object modeling. *IEEE Trans. Pattern Anal. Machine Intell.*, 14(2):157–173.

van Veen, J.A.J.C. and Werkhoven, P. 1996. Metamerisms in structure-from-motion perception. *Vision Research*, 36(14):2197–2210.

Woodham, R.J., Iwahori, Y., and Barman, R.A. 1991. Photometric stereo: Lambertian reflectance and light sources with unknown direction and strength. Technical Report 91-18, University of British Columbia, Laboratory for Computational Intelligence.

Zhang, Z. 1998. Image-based geometrically-correct photorealistic scene/object modeling (IBPhM): A review. In *Proc. 3rd Asian Conf. on Computer Vision*, pp. 340–349.

Zhao, C. and Mohr, R. 1996. Global three-dimensional surface reconstruction from occluding contours. *Computer Vision and Image Understanding*, 64(1):62–96.

Zitnick, C.L. and Webb, J.A. 1996. Multi-baseline stereo using surface extraction. Technical Report CMU-CS-96-196, Carnegie Mellon University, Pittsburgh, PA.