Approximate N-View Stereo

Kiriakos N. Kutulakos

Department of Computer Science & Department of Dermatology, Computer Studies Building,
University of Rochester, Rochester, NY 14627, USA
kyros@cs.rochester.edu

Abstract. This paper introduces a new multi-view reconstruction problem called *approximate N-view stereo*. The goal of this problem is to recover a one-parameter family of volumes that are increasingly tighter supersets of an unknown, arbitrarily-shaped 3D scene. By studying 3D shapes that reproduce the input photographs up to a special image transformation called a *shuffle transformation*, we prove that (1) these shapes can be organized hierarchically into nested supersets of the scene, and (2) they can be computed using a simple algorithm called *Approximate Space Carving* that is provably-correct for arbitrary discrete scenes (i.e., for unknown, arbitrarily-shaped Lambertian scenes that are defined by a finite set of voxels and are viewed from N arbitrarily-distributed viewpoints inside or around them). The approach is specifically designed to attack practical reconstruction problems, including (1) recovering shape from images with inaccurate calibration information, and (2) building coarse scene models from multiple views.

1 Introduction

The reconstruction of 3D objects and environments from photographs is becoming a key element in many applications that simulate physical interaction with the real world (e.g., [1]). Unfortunately, despite significant recent progress on the topic of N-view stereo [1–9], there are many practical reconstruction problems for which a general solution is beyond the current state of the art. Examples include (1) reconstructing an unknown scene from images with inaccurate calibration information [10, 11], (2) reconstructing a scene that is not perfectly stationary (e.g., a person that moved slightly between each snapshot [11]), and (3) recovering a coarse scene model either for efficiency reasons [12, 13], or because the scene's geometry is exceedingly complex (e.g., a tree full of leaves, viewed from a distance).

As a first step toward addressing this limitation, in this paper we consider a new multi-view reconstruction problem called *approximate N-view stereo*. The goal of this problem is to recover a one-parameter family of volumes that are increasingly tighter supersets of the true scene. Working from first principles, we show that each of the above reconstruction tasks can be thought of as instances of the approximate N-view stereo problem and, as such, they can be solved by recovering approximate, rather than exact, 3D shapes from multiple views. Moreover, we provide a detailed geometrical analysis of the approximate N-view stereo problem and describe a computational framework

The support of the National Science Foundation under Grant No. IIS-9875628, of Roche Laboratories, Inc., and of the Dermatology Foundation are gratefully acknowledged.

for provably solving it for discrete scenes in the general case (i.e., for an unknown, arbitrarily-shaped Lambertian scene that is viewed from N arbitrarily-distributed viewpoints inside or around it). Experimental results illustrating the applicability of our theoretical development are given for real scenes of complex geometry.

Our approach is motivated by the general question of how to recover 3D scene approximations from multiple views. We argue that any answer to this question should be evaluated according to the following criteria:

- **Direct recovery:** The approximation should be computable directly from the input images, i.e., without first computing an exact reconstruction.
- **Generality:** Computations should rely as little as possible on assumptions about the distribution of the input viewpoints (e.g., nearby views or minimal occlusion), about the scene's true shape, or about the existence of specific image features in the input views (e.g., edges, points, lines, contours, texture, or color).
- **Level-of-detail control:** It should be possible to control the degree to which the recovered shape approximates the shape of the true scene (to the extent allowed by the image data).
- **Shape determinism:** It should be possible to quantify the geometric relationship between the recovered approximation and the shape of the true scene.
- **Reconstructibility conditions:** It should possible to state the conditions under which the approximation process is valid and/or breaks down.
- **Robustness:** Approximate reconstruction should be possible even when ideal stereo conditions (e.g., good calibration, scene rigidity) are not satisfied.

Unfortunately, serious difficulties arise when attempting to fulfill the above criteria using existing approaches.

First, current theories for 3D shape approximation (e.g., scale space [14], mesh simplification [15], wavelet descriptions [16], and hierarchical volume representations [17]) define the notion of an "approximate 3D shape" in terms of the exact geometry of the shape being approximated. These theories are therefore not directly applicable when exact reconstructions are unavailable.

Second, even though recent approaches to N-view stereo have demonstrated good results even under very general conditions about the scene and the input viewpoints [5, 6, 8], they cannot be easily extended to achieve approximate reconstruction. Key to their success is the use of provably-correct algorithms to determine the scene points visible in each view. Little is known, however, about whether visibility determination is well defined and tractable when recovering approximate shapes from images. This is because even a small deviation from the shape of the true scene can imply dramatic changes in visibility [18], and because the input images will *not* be consistent, in general, with point visibilities defined by an approximate shape.

Third, existing stereo methods rely on the assumption that under ideal conditions, every point on a reconstructed 3D shape should project to points that can be matched across views. Unfortunately, this assumption cannot be used as a basis for designing approximate shape recovery algorithms because it breaks down in that case. Hence, even though recent particle- and mesh-based stereo techniques allow level-of-detail control [9, 11, 19, 20], their reliance on regularization criteria that enforce this assumption makes

it impossible to analyze their behavior during approximate reconstruction (e.g., existence of local minima, convergence properties, shape determinism).

Fourth, existing attempts to recover approximate shapes from multiple views rely exclusively on a two-step method that involves (1) reducing the resolution of the input views, and (2) recovering low-resolution shapes from the reduced images [12, 13]. In general, it is impossible to guarantee that low-resolution pixels spanning large depth or color discontinuities will be matched across views [12]. Approaches that rely on this method are therefore largely heuristic.

Fifth, with the exception of [11], no previous techniques have recognized the tight inter-dependence between the problem of approximate reconstruction and that of reconstruction in the presence of calibration errors. As a result, only partial studies exist for handling these problems.[1]

In order to overcome the limitations of existing approaches, our work is based on a simple idea: rather than define the notion of an "approximate 3D shape" directly in terms of 3D geometry, we define it *indirectly*, in terms of its appearance. To make this idea concrete, we use a class of image transformations suitable for describing a process of "controlled image approximation," and define approximate 3D shapes to be volumes in space that reproduce the input photographs up to a transformation in this class.

The two crucial questions one must address to exploit this idea are (1) how to relate these implicitly-defined approximate shapes to the geometry of the true scene, and (2) how to recover such shapes from images of arbitrarily-shaped scenes. Here we show that both questions can be answered with the help of a special class of image transformations called *shuffle transformations*. These transformations describe arbitrary bounded re-arrangements of points in an image and have a unique property—the views of the true scene are always related to the views of its supersets by a shuffle transformation. Moreover, we can use these transformations to arrange a scene's supersets into a one-parameter family of nested volumes, called the *Photo-Consistency Scale Space*, whose appearance converges to the input views and whose shapes provide increasingly tighter bounds on the true scene. Importantly, we show that shapes from this scale space can be computed using a simple, efficient and provably-correct volumetric algorithm that fulfills our stated evaluation criteria to a great degree. To our knowledge, this is the only algorithm with this property.

In the following we consider a scene to be an arbitrary, bounded, and opaque volume $\mathcal{V}$ that is viewed from arbitrary viewpoints in $\Re^3 - \mathcal{V}$. To simplify our exposition, we first focus on the case where volumes are not finite, i.e., they are open subsets of $\Re^3$ bounded by closed surfaces [21]; where images are functions of color or intensity defined over a continuous domain $[0, H] \times [0, W]$; where pixels are infinitesimally-small points; and where no pixel noise is present. We then consider finite volumes in Section 5 and discrete images in Section 6 and in the Appendix.

[1] For instance, the shape distortion analysis in [10] cannot be extended from 2 to N erroneously-calibrated views because *no* single 3D shape will be consistent, in general, with N such views—in such cases, only *approximate* shapes can be recovered. Similarly, the coarse-to-fine reconstruction method of Prock and Dyer [12] still requires accurate calibration.

Figure 1. Shuffle transformations. (a) A 1-shuffle corresponding to a piecewise-continuous image translation. (b) A 1-shuffle corresponding to a non-parametric image transformation; this example shows that r-shuffles can model the process of "ignoring" a subset of the pixels in an input image. (c),(d) A randomized 10-shuffle: the image in (d) was created by displacing every pixel in (c) to a randomly-selected position inside a 21x21 pixel window centered at the pixel. Note that (d) appears "blurred" even though no modification of colors or intensities has taken place.

2 Approximate N-View Stereo

A basic step in our method for recovering approximate 3D shapes is to apply in a novel way the principle of *transformation-invariant reconstruction* [22]: given a collection of input photographs, recover a 3D shape that is defined up to an *a priori* specified class of transformations. Below we apply this principle in an appearance based way by (1) defining the class of shuffle image transformations, and (2) defining approximate reconstruction as the problem of recovering a volume that reproduces the input photographs up to a shuffle transformation.

2.1 Shuffle Transformations

We use the term *shuffle transformation* to denote any image transformation (continuous or otherwise) that causes the bounded repositioning of pixels in an image. Shuffle transformations are defined implicitly, in terms of their effect on a source image:

Definition 1 (r-Shuffle) A 2D transformation $\mathcal{T} : I_1 \rightarrow I_2$ is called an *r-shuffle* if for every point in image I_2 we can find a point of identical color within a disk of radius r in I_1. The constant $r \geq 0$ is called the *dispersion radius* of $\mathcal{T}$.

Shuffle transformations affect only the arrangement of pixels in an image, not their actual colors or intensities (Figure 1a). Because these transformations encompass a wide range of distortions, algorithms that can recover shape from images known up to an r-shuffle are by definition invariant to bounded parametric, non-parametric, and statistically-defined image distortions (Figure 1b-d).

2.2 Reconstruction Using Photo-Consistency

Before we can make precise the approximate reconstruction problem, we need to relate the set of input views of an unknown scene to the 3D reconstructions they give rise to

in the ideal case, i.e., when an exact 3D reconstruction is sought. We use the photo-consistency theory of Kutulakos and Seitz [6] for this purpose, briefly summarized below.

The notion of photo-consistency is based on the idea that a reconstructed 3D shape should reproduce a scene's input photographs exactly if it is to be considered a valid geometric description of that scene. This leads to a geometric constraint satisfaction problem, where every input photograph can be thought of as a *constraint* that restricts the space of all possible 3D shapes to only those shapes that could have possibly produced that photograph. When many such photographs are available, each taken from a known position c_i, they define an equivalence class of 3D shape solutions called *photo-consistent shapes* whose views are identical to the input photographs when viewed from the photographs' viewpoints (Figure 2a):

Definition 2 (Photo-Consistent 3D Shape) An arbitrary finite and opaque volume $\mathcal{V}$ is *photo-consistent* if there is an assignment of radiances (colors) to every point on its surface such that $\mathcal{V}$'s projection along the known viewpoints $c_1, \ldots, c_N$ is identical to the corresponding input photographs.

Using this definition as a starting point, photo-consistency theory studies the reconstruction of photo-consistent shapes from N arbitrarily-distributed views that are taken from known positions in space. In particular, Kutulakos and Seitz proved the following:

Theorem 1 (Photo Hull Theorem) For every volume $\mathcal{V}$ that contains the true scene, there exists a unique shape, called the Photo Hull, that is the union of all photo-consistent subsets of $\mathcal{V}$. Moreover, this shape contains the true scene and is itself a photo-consistent shape.

The notion of the photo hull plays a key role in our approach for three reasons. First, it provides a direct mathematical link between a scene's appearance in N images and the reconstruction(s) these images can give rise to. Second, it defines in an algorithm-independent way the tightest possible superset of the true scene that we can ever recover from N photographs, in the absence of *a priori* scene information. This is because it is impossible to decide, based on the photographs alone, which photo-consistent subset of this maximal shape is the true scene. The notion of the photo hull is therefore especially important in order to evaluate the results of our approximate reconstruction approach. Third, it leads to a simple, efficient, and provably-correct volumetric algorithm for computing this shape—starting from an arbitrary superset $\mathcal{V}$ of the scene itself (e.g., an arbitrary "bounding box" that surrounds a physical 3D object), the algorithm iteratively "carves" voxels away from this superset until the carved volume converges to the photo hull. As such, photo-consistency theory leads to algorithms that satisfy both the Generality and Shape Determinism criteria posed in Section 1.

2.3 Approximation Using Shuffle Transformations

Unfortunately, despite its useful features, photo-consistency theory relies on precise knowledge of the input viewpoints and provides no mechanism for approximate shape

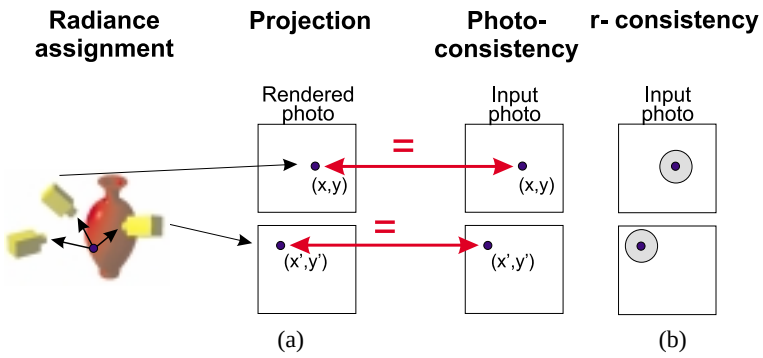

Figure 2. (a) Photo-consistent shapes. The color and intensity at the projection of every point on their surface must be identical to the input photographs. (b) r-consistent shapes. The color and intensity at the projection of every point on their surface must be identical to that of a pixel within a disk of radius r in the input photographs (shown in gray).

recovery. We therefore relax the definition of a photo-consistent 3D shape in a way that makes 3D shape approximation a mathematically tractable problem:

Definition 3 (r-Consistent 3D Shape) A volume $\mathcal{V}$ is *r-consistent* if for every input photograph I_i there exists an r-shuffle $\mathcal{T}_i : I_i \rightarrow I'_i$ that makes $\mathcal{V}$ photo-consistent with the photographs $I'_1, \ldots, I'_N$.

r-consistent shapes are the central concept in approximate N-view stereo. Intuitively, these shapes satisfy two seemingly-incompatible requirements. On one hand, a valid 3D scene approximation must be *globally consistent* with the input photographs, i.e., every point on its surface must conform to the textures, discontinuities, and occlusion relationships captured by the entire set of the input views. On the other, such an approximation cannot reproduce the input views exactly since it only approximates the scene itself.

When a point on an r-consistent shape is projected into a pair of photographs in which the point is visible, it induces an implicit correspondence between a pair of disks (Figures 2b, 3a). This correspondence can be interpreted in terms of a simple criterion for matching two or more sets of pixels:

Definition 4 (Color Equivalence of Pixel Sets) Two pixel sets are *color equivalent* if there is a pixel color that appears in both.

r-consistent shapes can therefore be thought of as integrating into a global, N-view stereo framework elements from recent non-parametric [23–25] and robust [26] approaches to image matching.

3 The Photo-Consistency Scale Space

Our definition of r-consistency leads directly to a hierarchy of 3D scene approximations in which the dispersion radius, r, controls how well the appearance of the 3D

approximation matches that of the true scene. To use this hierarchy as a framework for developing reconstruction algorithms, however, we must establish the 3D relationship between an r-consistent shape and the true scene.

We make this relationship precise with the following theorem which suggests that the dispersion radius can be thought of as a "scale" parameter that controls the detail in an r-consistent shape. Theorem 2 shows that the relationship between representations of the same scene at different scales is one of *containment*, i.e., r-consistent shapes at coarse scales are guaranteed to contain counterparts at finer scales (Figure 3a): [2]

Theorem 2 (Nesting Theorem) Let $\mathcal{V}$ be an arbitrary bounded superset of the scene:

1. 0-consistency is equivalent to photo-consistency;
2. r-consistency implies r_1-consistency for every $r_1 > r$;
3. every superset of the scene is r-consistent for some $r > 0$;
4. if $\mathcal{V}_r \subseteq \mathcal{V}$ is an r-consistent shape that contains the scene, then for every $r_1 \geq r$ we can find an r_1-consistent subset of $\mathcal{V}$; equivalently, for every $0 \leq r_2 \leq r$, we can find an r_2-consistent superset of the scene that is contained in $\mathcal{V}_r$.

The Nesting Theorem has three implications for approximate shape recovery. First, it suggests that the recovery of a photo-consistent or an r-consistent shape from photographs can always be formulated as a "coarse-to-fine" reconstruction process in which r-consistent shapes of increasingly smaller dispersion radius are recovered at each step, starting from an arbitrary initial volume that bounds the scene. Second, it shows that this process provides a way to reconstruct "controlled" scene approximations that (1) always bound the true scene, and (2) provide an increasingly accurate representation of both a scene's 3D shape and of its appearance. Third, it suggests that we can establish an explicit bound on a scene's 3D shape even from images that do *not* correspond to exact views of a rigid 3D scene. This is because the images of an r-consistent shape are defined only up to a shuffle transformation, and hence, any collection of r-shuffle-transformed views of a rigid scene are sufficient to determine such a shape.

4 Reconstruction Using Free-Space Queries

Our goal is to recover r-consistent shapes from multiple views of an arbitrary scene. To do this, we exploit the Nesting Theorem by repeatedly applying to an arbitrary superset of the scene an operation akin to "space carving:" given an r_1-consistent superset with $r_1 \geq r$, remove selected portions of this volume so that the resulting "carved" shape becomes r-consistent. In this section, we show how to perform this carving operation with the help of a provably-correct criterion for testing whether a portion of a non r-consistent volume is completely devoid of scene points.

The key observation we exploit to define this criterion is that even though r-consistency was defined in a purely appearance-based way, the r-consistency and non

[2] The reader should note that our approach differs from existing scale space theories (e.g., [14]) in two significant ways: (1) unlike previous formulations, the mapping between r-consistent shapes at different scales is *not* continuous in general, and (2) this mapping is defined in terms of the shapes' appearance rather than their geometry.

r-consistency of a shape provides *explicit* 3D geometric information about the underlying scene. In particular, let $\mathcal{V}$ be an arbitrary non r-consistent superset of the scene. By its definition, $\mathcal{V}$ must contain a point p on its surface such that (1) p is visible from a subset $\{c_1, \ldots, c_k\}$ of the input viewpoints, [3] and (2) the disks, $D_1, \ldots, D_k$, that are centered at p's projection in the corresponding photographs are not color equivalent. Let $\mathcal{R}_i, i = 1, \ldots, k$ be the interior of the conical volumes defined by c_i and D_i, respectively. We use the following theorem (Figures 3c,d):

Theorem 3 (Free-Space Query Theorem) (1) If the volume $\mathcal{R}_F = \bigcap_{i=1}^{k} \mathcal{R}_i$ is not occluded by $\mathcal{V} - \mathcal{R}_F$ from any of the viewpoints $c_i, i = 1, \ldots, k$, $\mathcal{R}_F \cap \mathcal{V}$ is free of scene points. (2) If $\mathcal{R}'_F$ is the subset of $\mathcal{R}_F$ that is unoccluded by $\mathcal{V} - \mathcal{R}_F$ from $c_i, i = 1, \ldots, k$, the volume $\mathcal{R}'_F \cap \mathcal{V}$ is free of scene points.

Theorem 3 gives a deterministic sufficiency criterion for "querying" the free space inside a non r-consistent volume by simply comparing disks around the projection of a single point on the volume's surface:

Corollary 1 (Free-Space Query Criterion) If the disks $D_1, \ldots, D_n$ are not color equivalent, there exists an identifiable volume $\mathcal{V}_F \subset \mathcal{V}$ that contains no scene points.

5 The Approximate Space Carving Algorithm

The Free Space Query criterion leads directly to a simple volumetric reconstruction algorithm that, given a dispersion radius r and an arbitrary superset $\mathcal{V}$ of the scene, iteratively removes portions of that volume until it becomes r-consistent.

In particular, the Free Space Theorem tells us that if there is a point p on $\mathcal{V}$'s surface that satisfies this criterion, the volume $\mathcal{V}' = \mathcal{V} - \mathcal{V}_F$ must still contain the scene. Furthermore, the Nesting Theorem guarantees the existence of an r-consistent superset of the scene in the volume $\mathcal{V}'$. Hence, if the scene consists of a finite set of points (e.g., voxels) and only the points in $\mathcal{V}_F$ are removed at each iteration, the carved volume will converge to an r-consistent shape. These considerations lead to the following algorithm for computing an r-consistent shape:

Approximate Space Carving Algorithm

Step 1: Initialize $\mathcal{V}$ to a superset of the scene.
Step 2: Repeat the following until no surface point p can be selected in Step 2b:
- **a.** Project p to all viewpoints $c_1, \ldots, c_j$ in which it is visible and let $D_1, \ldots, D_j$ be disks of radius r around p's projection in the corresponding photographs.
- **b.** Select p if no single pixel color appears in all disks.

Step 3: Let $\mathcal{V}_F$ be the largest volume in $\mathcal{V}$ that contains p, is fully visible in $c_1, \ldots, c_j$, and projects to the interior of $D_1, \ldots, D_j$, respectively.
Step 4: Set $\mathcal{V} = \mathcal{V} - \mathcal{V}_F$, and continue with Step 2.

[3] More formally, we consider $p \in \mathcal{V}$ to be *visible* to a set of cameras $\{c_1, \ldots, c_k\}$ if there exists a volume $\mathcal{V}' \subset \mathcal{V}$ around p whose occluders lie entirely in $\mathcal{V}'$. That is, for every $q \in \mathcal{V}'$, the open line segment qc_i does not intersect $\mathcal{V} - \mathcal{V}'$ for any camera c_i in the set.

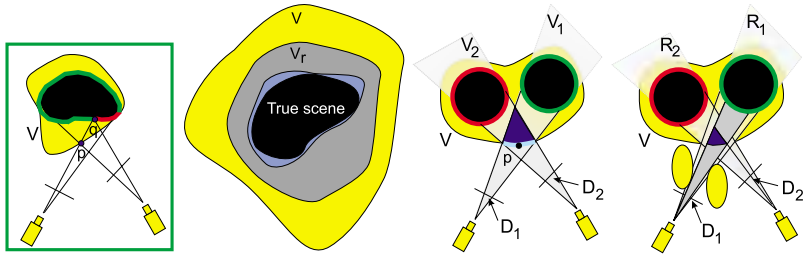

Figure 3. (a) Illustration of r-consistency for a shape $\mathcal{V}$ (shown in yellow) that contains the true scene. The scene's volume is shown in black. If the distance between the projection of p and q in the right-hand image is d, $\mathcal{V}$ will be r-consistent for every $r > d/2$. (b) Illustration of the Nesting Theorem. The blue volume represents the Photo Hull. (c)-(d) Illustrations of the Free-Space Theorem. (c) The intersection $\mathcal{R}_F = \mathcal{R}_1 \cap \mathcal{R}_2$ is unoccluded by $\mathcal{V}$; the volume $\mathcal{V}_F$, shown in dark blue, must be free of scene points—if it were not, at least one scene point would be visible to both cameras, forcing the color equivalence of the disks D_1, D_2. (d) The intersection $\mathcal{R}_1 \cap \mathcal{R}_2$ is occluded by $\mathcal{V}$; the volume $\mathcal{V}_F$ is restricted to the intersection of the visible portion of $\mathcal{R}_F$.

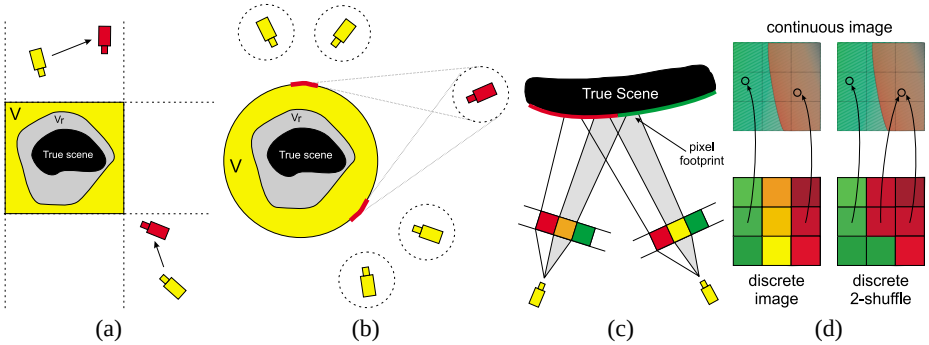

Figure 4. (a)-(b) Illustration of the Calibration Reconstructibility Theorem. (a) The initial volume $\mathcal{V}$, shown in yellow, is a bounding box of the scene. Yellow cameras indicate the true viewpoints, c_i. If the incorrect viewpoints $\hat{c}_i$, shown in red, do not cross the dotted lines, the volume $\mathcal{V}$ contains a reconstructible r-consistent shape. The theorem also provides an easily-computable reconstructibility test: if $\mathcal{V}$ itself is not r-consistent for any r, we cannot use the Approximate Space Carving Algorithm to reconstruct the scene. (b) Another application of the theorem. The circles around each camera are *a priori* bounds on the calibration error of the input views. An r-consistent shape can be recovered from the circular initial volume, $\mathcal{V}$, by allowing the red camera to contribute to the carving of $\mathcal{V}$ only around surface points outside the red regions, i.e., points whose visibility is not affected by calibration errors. (c)-(d) Handling discrete images. (c) The difference between the values of the corresponding red pixels, whose footprints contain no irradiance discontinuities, tends to zero with increasing image resolution. On the other hand, even though the middle pixels in the left and right image form a corresponding stereo pair, their actual values will be arbitrary mixtures of red and green and their similarity cannot be guaranteed for *any* image resolution. (d) Relating a discrete image to a 2-pixel-shuffle of its continuous counterpart. For every pixel in the discrete image whose footprint does not span a discontinuity, there is a point in the continuous image that has identical value and falls inside the footprint (circle). Hence, we obtain a 2-pixel-shuffle of the continuous image by replacing the value of every corrupted pixel in the discrete image by that of an adjacent pixel that does not lie on a discontinuity.

6 Applications of the Theory

Reconstruction in the Presence of Calibration Errors Because r-consistent reconstruction is invariant to shuffle transformations of the input images, it leads to algorithms that operate with predictable performance even when these views are not perfectly calibrated. More specifically, let $\mathcal{V}$ be an arbitrary superset of the true scene, let $c_1, \ldots, c_N$ be the viewpoints from which the input photographs were taken, and let $\hat{c_1}, \ldots, \hat{c_N}$ be incorrect estimates for these viewpoints, respectively.

Theorem 4 (Calibration Reconstructibility Theorem) *An r-consistent subset of $\mathcal{V}$ exists for some $r \geq 0$ if the following condition holds for all $i = 1, \ldots, N$:*

$$\text{Vis}_{\mathcal{V}}(c_i) = \text{Vis}_{\mathcal{V}}(\hat{c_i}), \tag{1}$$

where $\text{Vis}_{\mathcal{V}}(c)$ is the set of points on $\mathcal{V}$ that are visible from c.

Theorem 4 tells us that Approximate Space Carving can provably handle arbitrarily large calibration errors, as long as these errors do not affect the visibility of the initial volume given as input to the algorithm. This result is important because the conditions it sets are independent of the true scene—they only depend on the shape $\mathcal{V}$, which is known in advance (Figure 4).[4] In practice, even though large calibration errors may allow reconstruction of an r-consistent shape, they affect the minimum r for which this can happen. Hence, good information about a camera's calibration parameters is still needed to obtain r-consistent shapes that tightly bound the true scene. A key open question is whether it is possible to employ the approximate correspondences defined by an r-consistent shape to achieve *self-calibration*, i.e., improve camera calibration and recover even tighter bounds on the true scene [11, 27].

Reconstruction from Discrete Images In practice, pixels have a finite spatial extent and hence their color is an integral of the image irradiance over a non-zero solid angle. This makes it difficult to match individual pixels across views in a principled way, especially when the pixels span irradiance discontinuities (Figure 4c). Typical approaches use a threshold for pixel matching that is large enough to account for variations in the appearance of such "corrupted" pixels [8, 12]. Unfortunately, these variations depend on the radiance properties of the individual scene, they do not conform to simple models of image noise (e.g., Gaussian), and cannot be bounded for *any* finite image resolution. Unlike existing techniques, approximate N-view stereo puts forth an alternative approach: rather than attempting to model the appearance of corrupted pixels to match them across frames [3], we simply ignore these pixels altogether by recovering r-consistent shapes from the input views. Intuitively, this works because (1) a view of an r-consistent shape must agree with its corresponding discrete input image only up to a shuffle transformation, and (2) shuffle transformations are powerful enough to model the elimination of corrupted pixels from an input view by "pushing"

[4] It is possible to analyze in a similar way shape recovery when the scene moves between snapshots—reconstruction then involves computing r-consistent supersets of both the original *and* the new scene volume.

all discontinuities to pixel boundaries (Figures 1b and 4d). This behavior can be thought of as a generalization of the neiborhood-based method of Birchfield and Tomasi [28] where pixel (dis)similiarities are evaluated by comparing sets of intensities within their neighborhoods along an epipolar line. From a technical point of view, it is possible to establish a resolution criterion that is a sufficient condition for reconstructing r-consistent shapes from discrete images. This criterion formalizes the conditions under which the reasoning of Figure 4d is valid; it is omitted here due to lack of space. **Coarse-to-Fine Reconstruction** While the Approximate Space Carving Algorithm in Section 5 can be thought of as operating in a continuous domain, it can be easily converted into an algorithm that recovers volumetric scene models in a coarse-to-fine fashion [17]. The algorithm works by imposing a coarse discretization on the initial volume $\mathcal{V}$ and iteratively carving voxels away from it. The key question one must address is how to decide whether or not to carve a "coarse voxel" away. Figures 4c,d suggest that using lower-resolution images to achieve this [12, 13] will only aggravate the correspondence-finding process, and may lead to wrong reconstructions (i.e., coarse volumes that do not contain the scene). Instead, we apply the Free-Space Query Theorem: to decide whether or not to carve away a voxel v, we pick a dispersion radius r that is specific for that voxel and is large enough to guarantee that volume $\mathcal{V}_F$ contains v in its interior. Hence, carving coarse models simply requires establishing the color equivalence of disks in each image that are of the appropriate size. Moreover, the Nesting Theorem tells us that this approach not only guarantees that coarse reconstructions contain the scene; it also ensures that they can be used to derive higher-resolution ones by simply removing high-resolution voxels.

7 Experimental Results

To demonstrate the applicability of our approach, we performed several experiments with real image sequences. In all examples, we used a coarse-to-fine volumetric implementation of the Approximate Space Carving Algorithm as outlined in Section 6. No background subtraction was performed on any of the input sequences. We relied on the algorithm in the Appendix to test the color equivalence of disks across views. Voxels were assigned a color whose RGB values led to **component success** in this test.

Coarse-to-fine reconstruction: We first ran the Approximate Space Carving Algorithm on 16 images of a gargoyle sculpture, using a small threshold of 4% for the color equivalence test. 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. The sub-pixel calibration error in this sequence enabled a high-resolution reconstruction that was carved out of a 256^3 voxel volume. The maximum dispersion radius was approximately two pixels. Figure 5 also shows coarse reconstructions of the same scene carved out of 128^3, 64^3, and 32^3 voxel volumes. The only parameter changed to obtain these volumes was the voxel size. Note that the reconstructions preserve the

object's overall shape and appearance despite the large dispersion radii associated with them (over 40 pixels for the 32^3 reconstruction).

Invariance under bounded image distortions: To test the algorithm's ability to reconstruct 3D shapes from "approximate" images, we ran the approximate space carving algorithm on artificially-modified versions of the gargoyle sequence. In the first such experiment, we shifted each input image by a random 2D translation of maximum length d along each axis. These modifications result in a maximum dispersion error of $2\sqrt{2}d$ for corresponding pixels. The modified images can be thought of either (1) as erroneously-calibrated input images, whose viewpoint contains a translation error parallel to the image plane, or (2) as snapshots of a moving object taken at different time instants. Figure 5 shows a 128^3 reconstruction obtained for $d = 3$ in which the approximate space carving algorithm was applied with exactly the same parameters as those used for the original images. Despite the large dispersion errors, the recovered shape is almost identical to that of the "error-free" case, even though the algorithm was run without modifying any of the input parameters. This is precisely as predicted by our theory: since the dispersion radius associated with each voxel is larger than 8.5 pixels, the recovered r-consistent shape is guaranteed to be invariant to such errors. A 64^3 reconstruction for $d = 10$ is also shown in the figure, corresponding to dispersion errors of over 22 pixels. In a second experiment, we individually shifted every pixel within a 21×21-pixel window, for every image of the gargoyle sequence (Figure 1d shows one image from the modified sequence). Figure 5 shows a 64^3 reconstruction from this sequence that was obtained by running the approximate space carving algorithm with exactly the same parameters as those used to obtained the 64^3 reconstruction for the original sequence. These results suggest that our approach can handle very large errors and image distortions without requiring any assumptions about the scene's shape, its appearance, or the input viewpoints, and without having to control any additional parameters to achieve this.

Reconstruction from images with mixed pixels: In order to test our algorithm's performance on sequences where many of the image pixels span color and intensity discontinuities, we ran the algorithm on 30 calibrated images of four cactus plants. The complex geometry of the cacti creates a difficult stereo problem in which the frequent discontinuities and discretization effects become significant. Our results show that the dispersion radius implied by a 128^3 volume results in a good reconstruction of the scene using a low 7% RGB component error. Despite the presence of a significant number of "mixed" pixels in the input data, which make stereo correspondences difficult to establish, the reconstruction does not show evidence of "over-carving" (Figure 5).

Reconstruction from images with calibration errors: To test invariance against calibration errors, we ran the approximate space carving algorithm on 45 images of an African violet. While calibration information for each of these views was available, it was not extremely accurate, resulting in correspondence errors of between 1 and 3 pixels

in the input views.[5] Figure 6 shows results of the reconstructed 3D shape, achieved by reconstructing a 128^3 volume from the input views. This result suggests that the dispersion radius implied by the 128^3 reconstruction was sufficient to handle inaccuracies in the calibration of each view. An important future direction for our work will be to examine how to use this approximately-reconstructed shape for self-calibration, i.e., for further improving the calibration information of each view and, hence, increasing the maximum attainable resolution of the recovered shape.

Comparison to standard space carving: In a final set of experiments, we compared our approach to a volumetric reconstruction algorithm that does not incorporate an error-tolerance model. To achieve this, we applied an implementation of the space carving algorithm [6] to the same sequences of images used in the above experiments, with exactly the same parameters. To obtain a coarse scene reconstruction of the gargoyle, we ran the space carving algorithm on a volumetric grid of the desired resolution (i.e., a cube of 128^3 voxels for a 128^3 reconstruction). In all of our coarse reconstruction runs at $128^3, 64^3, 32^3$ resolutions, the space carving algorithm completely failed to reconstruct the scene; instead, the entire set of voxels in the starting volume was carved away resulting in empty reconstructions. Intuitively, since larger voxels project to larger regions in the input views, their projection spans pixels with significant intensity variation, invalidating the voxel consistency criterion employed by that algorithm. Our approximate reconstruction technique, on the other hand, not only produces valid coarse reconstructions, but does so even when the input views are distorted significantly.

We next ran the standard space carving algorithm on the cactus sequence. In this case, even a reconstruction of 256^3, where every voxel projects to approximately one pixel, led to over-carving in some parts of the scene (Figure 7). Attempts to reconstruct lower-resolution volumes with the algorithm led to almost complete carving of the input volume. A similar behavior was exhibited with the violet sequence. In this case, calibration errors in the input views led to significant over-carving even for the highest-possible volume resolution of 256^3, where every voxel projects to about one pixel. This suggests that our approximate reconstruction algorithm does not suffer from the original algorithm's sensitivity to even relatively small calibration errors.

8 Concluding Remarks

Despite our encouraging results, several important questions still remain unanswered. These include (1) how to determine the minimum dispersion radius that leads to valid scene reconstructions, (2) how to apply our analysis to the problem of joint reconstruction/self-calibration from N views, and (3) how to develop adaptive, multi-resolution reconstruction methods in which different parts of the same scene are approximated to different degrees.

While invariance under shuffle transformations leads to increased robustness against calibration errors and discretization effects, we believe that shuffle-invariant reconstruction is only a first step toward a general theory of approximate shape recovery. This is

[5] This error range was established *a posteriori* from images of a test pattern that was viewed from approximately the same camera positions but that was not used for calibration.

because calibration errors and image distortions are "structured" to some extent, and rarely correspond to truly arbitrary bounded pixel repositionings. As such, our current investigation can be thought of as treating a worst-case formulation of the approximate N-view stereo problem. The question of how to limit the class of acceptable image distortions without compromising robustness in the shape recovery process is a key aspect of our ongoing research.

Appendix: Implementation of the Color Equivalence Test Step 2b of the Approximate Space Carving Algorithm requires determining whether two or more sets of n pixels share a common color. This requires $O(n^2)$ pixel comparisons for two n-pixel disks (i.e., examining every possible pixel pair (q_1, q_2) with $q_1 \in D_1$ and $q_2 \in D_2$). To improve efficiency, we use the following algorithm, which requires only $O(kn \log n)$ pixel comparisons for k disks:

Step 0 Return **success** if and only if Steps 1-3 return **component success** for every color component $C = \{R, G, B\}$. In this case, assign color (μ_R, μ_G, μ_B) to the associated voxel.

Step 1 Let $R_i, i = 1, \ldots, k$ be the array of values for component C in the pixels of D_i.

Step 2 Sort R_1 and repeat for each disk $D_i, i = 2, \ldots, k$:

a. sort R_i;

b. for every value $R_1[j]$, find its closest value, R_i^j, in R_i.

Step 3 For every value $R_1[j]$, compute the standard deviation, σ_j, of the values in the set $\{R_1[j], R_2^j, \ldots, R_k^j\}$.

Step 4 If $\min_j \sigma_j < \tau$, where τ is a threshold, return the mean, μ_C, of the component values along wtih **component success**; otherwise, return **failure**.

This algorithm is equivalent to the Color Equivalence Test when D_1 degenerates to a pixel and weakens with increasing disk size. The threshold τ is chosen by assuming Gaussian noise of known standard deviation σ^0 for every pixel component. Specifically, the sample variance, σ_j^2, follows a χ^2 distribution with $k - 1$ degrees of freedom in the case of **component success**. The threshold can therefore be chosen to ensure that **component success** is returned with high probability when the disks share a common value for component C [29].

References

1. P. J. Narayanan, P. W. Rander, and T. Kanade, "Constructing virtual worlds using dense stereo," in *Proc. Int. Conf. on Computer Vision*, pp. 3–10, 1998.
2. S. B. Kang and R. Szeliski, "3-D scene data recovery using omnidirectional multibaseline stereo," in *Proc. Computer Vision and Pattern Recognition Conf.*, pp. 364–370, 1996.
3. R. Szeliski and P. Golland, "Stereo matching with transparency and matting," in *Proc. 6th Int. Conf. on Computer Vision*, pp. 517–524, 1998.
4. S. Roy and I. J. Cox, "A maximum-flow formulation of the N-camera stereo correspondence problem," in *Proc. 6th Int. Conf. on Computer Vision*, pp. 492–499, 1998.
5. O. Faugeras and R. Keriven, "Complete dense stereovision using level set methods," in *Proc. 5th European Conf. on Computer Vision*, pp. 379–393, 1998.

6. K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," in *Proc. 7th Int. Conf. Computer Vision*, pp. 307–314, 1999.
7. R. T. Collins, "A space-sweep approach to true multi-image matching," in *Proc. Computer Vision and Pattern Recognition Conf.*, pp. 358–363, 1996.
8. S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by voxel coloring," in *Proc. Computer Vision and Pattern Recognition Conf.*, pp. 1067–1073, 1997.
9. P. Fua and Y. G. Leclerc, "Object-centered surface reconstruction: Combining multi-image stereo and shading," *Int. J. Computer Vision*, vol. 16, pp. 35–56, 1995.
10. G. Baratoff and Y. Aloimonos, "Changes in surface convexity and topology caused by distortions of stereoscopic visual space," in *Proc. 5th European Conf. on Computer Vision*, pp. 226–240, 1998.
11. P. Fua and Y. G. Leclerc, "Registration without correspondences," in *Proc. Computer Vision and Pattern Recognition Conf.*, pp. 121–128, 1994.
12. A. C. Prock and C. R. Dyer, "Towards real-time voxel coloring," in *Proc. Image Understanding Workshop*, pp. 315–321, 1998.
13. R. Szeliski, "Rapid octree construction from image sequences," *CVGIP: Image Understanding*, vol. 58, no. 1, pp. 23–32, 1993.
14. T. Lindeberg, *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.
15. J. Popovic and H. Hoppe, "Progressive simplicial complexes," in *Proc. SIGGRAPH'97*, pp. 189–198, 1997.
16. E. J. Stollnitz, T. D. DeRose, and D. H. Salesin, *Wavelets for Computer Graphics*. Morgan Kaufmann Publishers, 1996.
17. H. Samet, *The design and analysis of spatial data structures*. Reading, MA: Addison-Wesley, 1990.
18. J. J. Koenderink, *Solid Shape*. MIT Press, 1990.
19. P. Fua, "From multiple stereo views to multiple 3-d surfaces," *Int. J. Computer Vision*, vol. 24, no. 1, pp. 19–35, 1997.
20. R. Szeliski and H.-Y. Shum, "Motion estimation with quadtree splines," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, no. 12, pp. 1199–1210, 1996.
21. M. A. Armstrong, *Basic Topology*. Springer-Verlag, 1983.
22. J. L. Mundy and A. Zisserman, eds., *Geometric Invariance in Computer Vision*. MIT Press, 1992.
23. D. N. Bhat and S. K. Nayar, "Ordinal measures for visual correspondence," in *Proc. Computer Vision and Pattern Recognition Conf.*, pp. 351–357, 1996.
24. J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image indexing using color correlograms," in *Proc. Computer Vision and Pattern Recognition Conf.*, pp. 762–768, 1997.
25. R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. 3rd European Conf. on Computer Vision*, pp. 151–158, 1994.
26. P. Meer, D. Mintz, A. Rosenfeld, and D. Y. Kin, "Robust regression methods for computer vision: A review," *Int. J. Computer Vision*, vol. 6, no. 1, pp. 59–70, 1991.
27. S. J. Maybank and O. D. Faugeras, "A theory of self-calibration of a moving camera," *Int. J. Computer Vision*, vol. 8, pp. 123–152, 1992.
28. S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 4, pp. 401–406, 1998.
29. H. Cramér, *Mathematical Methods of Statistics*. Princeton University Press, 1946.

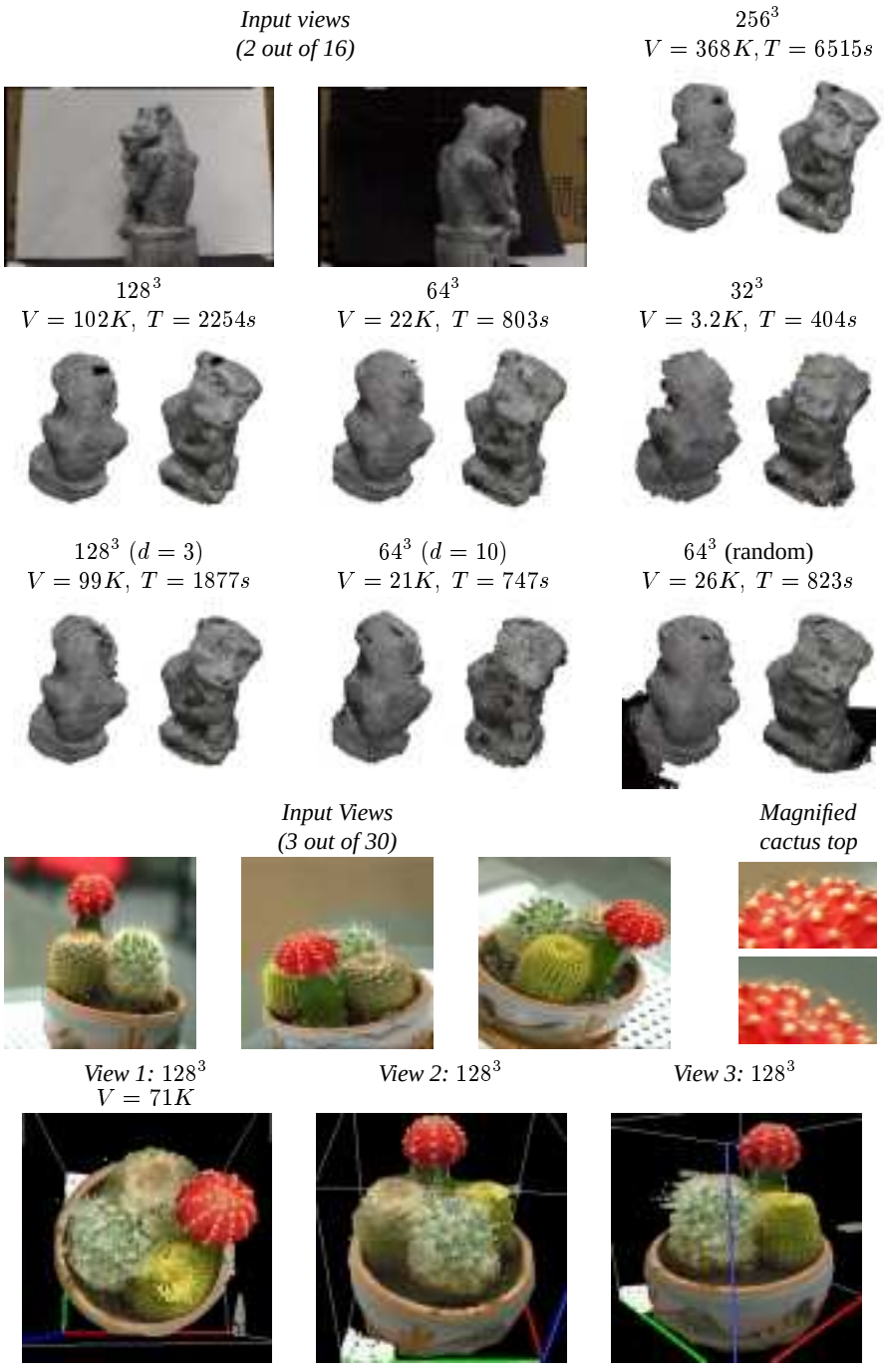

Figure 5. Experimental results. Also shown are the number of surface voxels, V, as well as the computation time, T, in seconds.

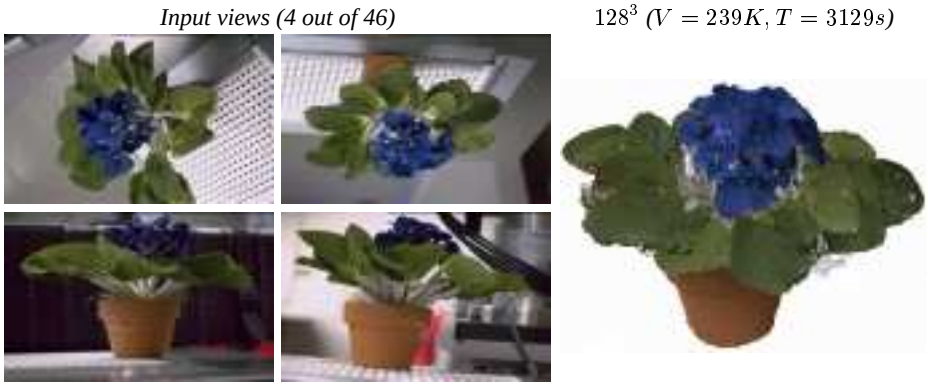

Figure 6. Reconstruction in the presence of calibration errors. Four out of 46 images of the sequence are shown. A view of the reconstructed 3D shape is shown on the right, obtained after full convergence of the approximate space carving algorithm.

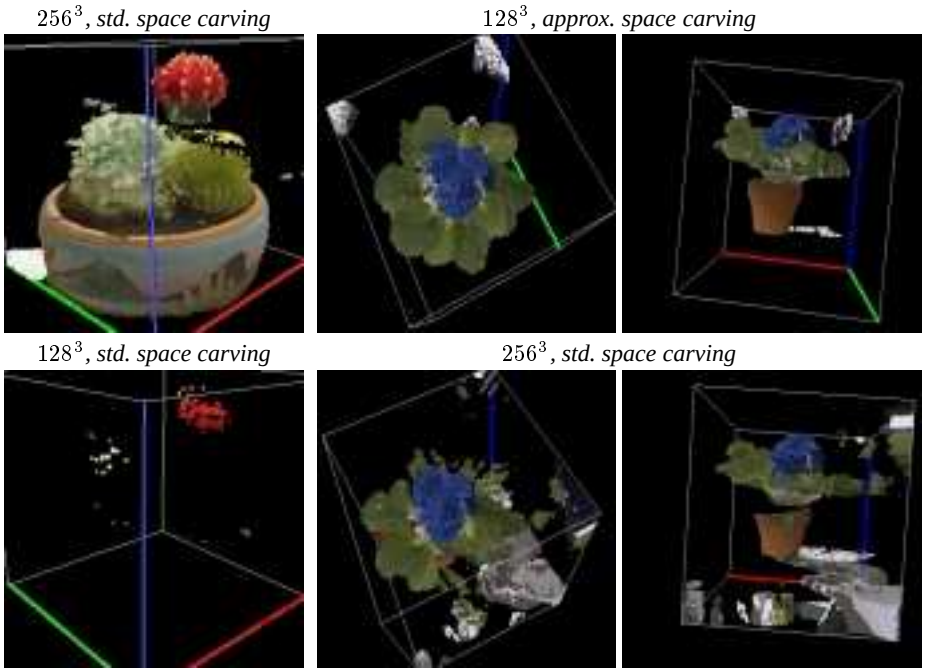

Figure 7. Comparing reconstructions computed with the standard and approximate space carving algorithm. *Left column:* Views of the reconstructed cacti. Over-carving is noticeable in the upper-right part of the recovered volume. *Right columns, top row:* More views of the reconstruction shown in Figure 6. *Right columns, bottom row:* Volume generated by standard space carving before the algorithm reached complete convergence. Note that even at this stage, significant over-carving is evident throughout the volume; the final reconstruction contains even fewer voxels, because over-carved regions cause errors in the algorithm's visibility computations.