

A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms

Steven M. Seitz Brian Curless
University of Washington

James Diebel
Stanford University

Daniel Scharstein
Middlebury College

Richard Szeliski
Microsoft Research

Abstract

This paper presents a quantitative comparison of several multi-view stereo reconstruction algorithms. Until now, the lack of suitable calibrated multi-view image datasets with known ground truth (3D shape models) has prevented such direct comparisons. In this paper, we first survey multi-view stereo algorithms and compare them qualitatively using a taxonomy that differentiates their key properties. We then describe our process for acquiring and calibrating multi-view image datasets with high-accuracy ground truth and introduce our evaluation methodology. Finally, we present the results of our quantitative comparison of state-of-the-art multi-view stereo reconstruction algorithms on six benchmark datasets. The datasets, evaluation details, and instructions for submitting new models are available online at <http://vision.middlebury.edu/mview>.

1. Introduction

The goal of multi-view stereo is to reconstruct a complete 3D object model from a collection of images taken from known camera viewpoints. Over the last few years, a number of high-quality algorithms have been developed, and the state of the art is improving rapidly. Unfortunately, the lack of benchmark datasets makes it difficult to quantitatively compare the performance of these algorithms and to therefore focus research on the most needed areas of development.

The situation in binocular stereo, where the goal is to produce a dense depth map from a pair of images, was until recently similar. Here, however, a database of images with ground-truth results has made the comparison of algorithms possible and hence stimulated an even faster increase in algorithm performance [1].

In this paper, we aim to rectify this imbalance by providing, for the first time, a collection of high-quality calibrated multi-view stereo images registered with ground-truth 3D models and an evaluation methodology for comparing multi-view algorithms.

Our paper's contributions include a taxonomy of multi-view stereo reconstruction algorithms inspired by [1] (Sec-

tion 2), the acquisition and dissemination of a set of calibrated multi-view image datasets with high-accuracy ground-truth 3D surface models (Section 3), an evaluation methodology that measures reconstruction accuracy and completeness (Section 4), and a quantitative evaluation of some of the currently best-performing algorithms (Section 5). While the current evaluation only includes methods whose authors were able to provide us their results by CVPR final submission time, our datasets and evaluation results are publicly available [2] and open to the general community. We plan to regularly update the results, and publish a more comprehensive comparative evaluation as a full-length journal publication.

We limit the scope of this paper to algorithms that reconstruct dense object models from calibrated views. Our evaluation therefore does not include traditional binocular, trinocular, and multi-baseline stereo methods, which seek to reconstruct a single depth map, or structure-from-motion and sparse stereo methods that compute a sparse set of feature points. Furthermore, we restrict the current evaluation to objects that are nearly Lambertian, which is assumed by most algorithms. However, we also captured and plan to provide datasets of specular scenes and plan to extend our study to include such scenes in the future.

This paper is not the first to survey multi-view stereo algorithms; we refer readers to nice surveys by Dyer [3] and Slabaugh *et al.* [4] of algorithms up to 2001. However, the state of the art has changed dramatically in the last five years, warranting a new overview of the field. In addition, this paper provides the first *quantitative* evaluation of a broad range of multi-view stereo algorithms.

2. A multi-view stereo taxonomy

One of the challenges in comparing and evaluating multi-view stereo algorithms is that existing techniques vary significantly in their underlying assumptions, operating ranges, and behavior. Similar in spirit to the binocular stereo taxonomy [1], we categorize existing methods according to six fundamental properties that differentiate the major algorithms: the *scene representation*, *photo-consistency measure*, *visibility model*, *shape prior*, *reconstruction algorithm*, and *initialization requirements*.

2.1. Scene representation

The geometry of an object or scene can be represented in numerous ways; the vast majority of multi-view algorithms use voxels, level-sets, polygon meshes, or depth maps. While some algorithms adopt a single representation, others employ different representations for various steps in the reconstruction pipeline. In this section we give a very brief overview of these representations and in Section 2.5 we discuss how they are used in the reconstruction process.

Many techniques represent geometry on a regularly sampled 3D grid (volume), either as a discrete occupancy function (e.g., voxels [5–19]), or as a function encoding distance to the closest surface (e.g., level-sets [20–26]). 3D grids are popular for their simplicity, uniformity, and ability to approximate any surface.

Polygon meshes represent a surface as a set of connected planar facets. They are efficient to store and render and are therefore a popular output format for multi-view algorithms. Meshes are also particularly well-suited for visibility computations and are also used as the central representation in some algorithms [27–32].

Some methods represent the scene as a set of depth maps, one for each input view [33–38]. This multi-depth-map representation avoids resampling the geometry on a 3D domain, and the 2D representation is convenient particularly for smaller datasets. An alternative is to define the depth maps relative to scene surfaces to form a *relief surface* [39, 40].

2.2. Photo-consistency measure

Numerous measures have been proposed for evaluating the visual compatibility of a reconstruction with a set of input images. The vast majority of these measures operate by comparing pixels in one image to pixels in other images to see how well they correlate. For this reason, they are often called *photo-consistency* measures [11]. The choice of measure is not necessarily intrinsic to a particular algorithm—it is often possible to take a measure from one method and substitute it in another. We categorize photo-consistency measures based on whether they are defined in *scene space* or *image space* [22].

Scene space measures work by taking a point, patch, or volume of geometry, projecting it into the input images, and evaluating the amount of mutual agreement between those projections. A simple measure of agreement is the variance of the projected pixels in the input images [8, 11]. Other methods compare images two at a time, and use window-matching metrics such as sum of squared differences or normalized cross correlation [20, 23, 31]. An interesting feature of scene-space window-based methods is that the current estimate of the geometry can inform the size and shape of the window [20]. A number of other photo-consistency

measures have been proposed to provide robustness to small shifts and other effects [12, 18].

Image space methods use an estimate of scene geometry to warp an image from one viewpoint to predict a different view. Comparing the predicted and measured images yields a photo-consistency measure known as *prediction error* [26, 41]. While prediction error is conceptually very similar to scene space measures, an important difference is the domain of integration. Scene space error functions are integrated over a surface and thus often tend to prefer smaller surfaces, whereas prediction error is integrated over the set of images of a scene and thus ascribe more weight to parts of the scene that appear frequently or occupy a large image area.

While most stereo algorithms have traditionally assumed approximately view-independent intensities, i.e., Lambertian scenes, a number of new photo-consistency metrics have been devised that seek to model more general reflection functions (BRDFs) [15–17, 22, 23, 32]. Some methods also utilize silhouettes [27, 30, 31] or shadows [17, 42].

2.3. Visibility model

Visibility models specify which views to consider when evaluating photo-consistency measures. Because scene visibility can change dramatically with viewpoint, almost all modern multi-view stereo algorithms account for occlusions in some way or another. Early algorithms that did not model visibility [6, 27, 43] have trouble scaling to large distributions of viewpoints. Techniques for handling visibility include *geometric*, *quasi-geometric*, and *outlier-based* approaches.

Geometric techniques seek to explicitly model the image formation process and the shape of the scene to determine which scene structures are visible in which images. A common approach in surface evolution approaches is to use the current estimate of the geometry to predict visibility for every point on that surface [5, 11, 12, 19, 20, 29, 30, 40]. Furthermore, if the surface evolution begins with a surface that encloses the scene volume and evolves by carving away that volume, this visibility approach can be shown to be *conservative* [11, 18]; i.e., the set of cameras for which a scene point is predicted to be visible is a subset of the set of cameras in which that point is truly visible.

Visibility computations can be simplified by constraining the allowable distribution of camera viewpoints. If the scene lies outside the convex hull of the camera centers, the occlusion ordering of points in the scene is same for all cameras [8], enabling a number of more efficient algorithms [8, 10, 13, 35, 44].

Quasi-geometric techniques use approximate geometric reasoning to infer visibility relationships. For example, a popular heuristic for minimizing the effects of occlusions is to limit the photo-consistency analysis to clusters of nearby cameras [31, 45]. This approach is often used in combi-

nation with other forms of geometric reasoning to avoid oblique views and to minimize computations [5, 11, 26]. Another common quasi-geometric technique is to use a rough estimate of the surface such as the visual hull [46] to guess visibility for neighboring points [19, 47, 48].

The third type of method is to avoid explicit geometric reasoning and instead treat occlusions as *outliers* [31, 34, 37, 38]. Especially in cases where scene points are visible more often than they are occluded, simple outlier rejection techniques [49] can be used to select the good views. A heuristic often used in tandem with outlier rejection is to avoid comparing views that are far apart, thereby increasing the likely percentage of inliers [31, 34, 37, 38].

2.4. Shape prior

Photo-consistency measures alone are not always sufficient to recover precise geometry, particularly in low-textured scene regions [11, 50]. It can therefore be helpful to impose *shape priors* that bias the reconstruction to have desired characteristics. While priors are essential for binocular stereo, they play a less important role in multi-view stereo where the constraints from many views are stronger.

Techniques that minimize scene-based photo-consistency measures naturally seek *minimal surfaces* with small overall surface area. This bias is what enables many level-set algorithms to converge from a gross initial shape [20]. The preference for minimal surfaces can also result in a tendency to smooth over points of high curvature (see [51, 52] for ways to address this problem). Recent approaches based on volumetric min-cut [19, 47] also have a bias for minimum surfaces. A number of mesh-based algorithms incorporate terms that cause triangles to shrink [29, 31] or prefer reference shapes such as a sphere or a plane [27].

Many methods based on voxel coloring and space carving [5, 8, 9, 11, 12, 16, 18, 53] instead prefer *maximal surfaces*. Since these methods operate by removing voxels only when they are not photo-consistent, they produce the largest photo-consistent scene reconstruction, known as the “photo hull.” Because they do not assume that the surface is smooth, these techniques are good at reconstructing high curvature or thin structures. However, the surface tends to bulge out in regions of low surface texture [8, 11].

Rather than impose global priors on the overall size of the surface, other methods employ shape priors that encourage local smoothness. Approaches that represent the scene with depth maps typically optimize an *image-based* smoothness term [33–37, 45] that seeks to give neighboring pixels the same depth value. This kind of prior fits nicely into a 2D Markov Random Field (MRF) framework, and can therefore take advantage of efficient MRF solvers [35]. A disadvantage is that there is a bias toward fronto-parallel surfaces. This bias can be avoided by enforcing surface-

based priors, as in [27, 29–32, 40, 47, 48].

2.5. Reconstruction algorithm

Multi-view stereo algorithms can be roughly categorized into four classes.

The first class operates by first computing a cost function on a 3D volume, and then extracting a surface from this volume. A simple example of this approach is the voxel coloring algorithm and its variants [8, 17], which make a single sweep through the volume, computing costs and reconstructing voxels with costs below a threshold in the same pass (note that [13] avoids the need for a threshold). Other algorithms differ in the definition of the cost function and the surface extraction method. A number of methods define a volumetric MRF and use max-flow [6, 19, 47, 48] or multi-way graph cut [35] to extract an optimal surface.

The second class of techniques works by iteratively evolving a surface to decrease or minimize a cost function. This class includes methods based on voxels, level sets, and surface meshes. Space carving [5, 11] and its variants [9, 11, 12, 14, 18, 40, 53] progressively remove inconsistent voxels from an initial volume. Other variants of this approach enable adding as well as deleting voxels to minimize an energy function [15, 54]. Level-set techniques minimize a set of partial differential equations defined on a volume. Like space carving methods, level-set methods typically start from a large initial volume and shrink inward; unlike most space carving methods, however, they can also locally expand if needed to minimize an energy function. Other approaches represent the scene as an evolving mesh [27–32] that moves as a function of internal and external forces.

In the third class are image-space methods that compute a set of depth maps. To ensure a single consistent 3D scene interpretation, these methods enforce consistency constraints between depth maps [33, 35–37], or merge the set of depth maps into a 3D scene as a post process [45].

The final class consists of algorithms that first extract and match a set of feature points and then fit a surface to the reconstructed features [55–58].

2.6. Initialization requirements

In addition to a set of calibrated images, all multi-view stereo algorithms assume or require as input some information about the geometric extent of the object or scene being reconstructed. Providing some constraints on scene geometry is in fact necessary to rule out trivial shapes, such as a different postcard placed in front of each camera lens.

Many algorithms require only a rough bounding box or volume (e.g., space carving variants [8, 9, 11, 12, 14, 18, 40, 53] and level-set algorithms [20–26]). Some algorithms require a foreground/background segmentation (i.e., silhouette) for each input image and reconstruct a visual

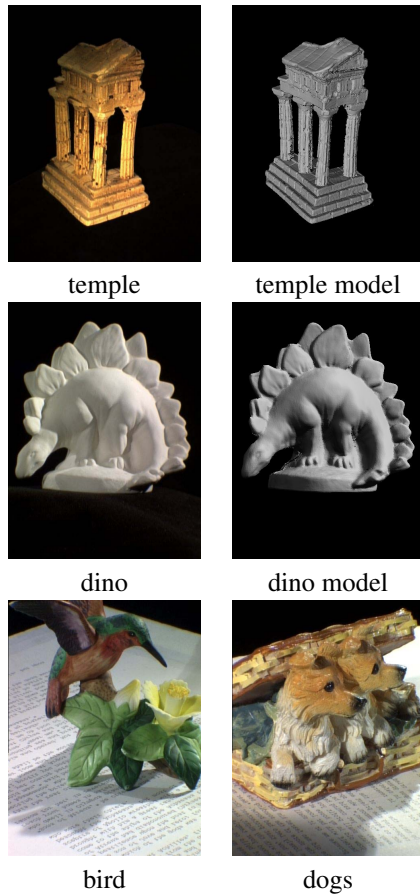


Figure 1. Multi-view datasets with laser-scanned 3D models.

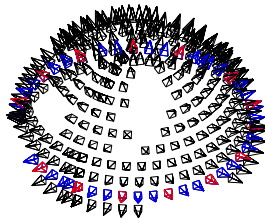


Figure 2. The 317 camera positions and orientations for the temple dataset. The gaps are due to shadows. The 47 cameras corresponding to the ring dataset are shown in blue and red, and the 16 sparse ring cameras only in red.

hull [46] that serves as an initial estimate of scene geometry [5, 19, 31, 47, 48].

Image-space algorithms [33, 35–37] typically enforce constraints on the allowable range of disparity or depth values, thereby constraining scene geometry to lie within a near and far depth plane for each camera viewpoint.

3. Multi-view data sets

To enable a quantitative evaluation of multi-view stereo reconstruction algorithms, we collected several calibrated

multi-view image sets and corresponding ground truth 3D mesh models. Similar data are available for surface lightfield studies [59, 60]; we have followed similar procedures for acquiring the images and models and for registering them to one another (although we add a step to automatically refine the alignment of the ground truth to the image sets based on minimizing photo-consistency). The surface lightfield data sets themselves are not, however, suitable for this evaluation due to the highly specular nature of the objects selected for those studies. We note that a number of other high quality multi-view datasets are publicly available (without registered ground truth models), and we provide links to many of these through our web site.

The target objects for this study were selected to have a variety of characteristics that are challenging for typical multi-view stereo reconstruction algorithms. We sought objects that broadly sample the space of these characteristics by including both sharp and smooth features, complex topologies, strong concavities, and both strongly and weakly textured surfaces (see Figure 1).

The images were captured using the Stanford spherical gantry, a robotic arm that can be positioned on a one-meter radius sphere to an accuracy of approximately 0.01 degrees. Images were captured using a CCD camera with a resolution of 640×480 pixels attached to the tip of the gantry arm. At this resolution, a pixel in the image spans roughly $0.25mm$ on the surface of the object (the temple object is $10cm \times 16cm \times 8cm$, and the dino is $7cm \times 9cm \times 7cm$).

The system was calibrated by imaging a planar calibration grid from 68 viewpoints over the hemisphere and using [61] to compute intrinsic and extrinsic parameters. From these parameters, we computed the camera's translational and rotational offset relative to the tip of the gantry arm, enabling us to determine the camera's position and orientation as a function of any desired arm position.

The target object sits on a stationary platform near the center of the gantry sphere and is lit by three external spotlights. Because the gantry casts shadows on the object in certain viewpoints, we double-covered the hemisphere with two different arm configurations, capturing a total of 790 images. After shadowed images were manually removed, we obtained roughly 80% coverage of the sphere. From the resulting images, we created three datasets for each object, corresponding to a full hemisphere, a single ring around the object, and a sparsely sampled ring (Figure 2).

The reference 3D model was captured using a Cyberware Model 15 laser stripe scanner. This unit has a single-scan resolution of $0.25mm$ and an accuracy of $0.05mm$ to $0.2mm$, depending on the surface characteristics and the viewing angle. For each object, roughly 200 individual scans were captured, aligned and merged on a $0.25mm$ grid, with the resulting mesh extracted with sub-voxel precision [62]; the accuracy of the combined scans is appreciably

greater than the individual scans. The procedure also produces per-vertex confidence information, which we use in the evaluation procedure.

The reference models were aligned to their image sets using an iterative optimization approach that minimizes a photo-consistency function between the reference mesh and the images. The alignment parameters consist of a translation, rotation, and uniform scale. The scale factor was introduced to compensate for small differences in calibration between the laser scanner and each image set. The photo-consistency function for each vertex of the mesh is the variance of the color of all rays impinging on that vertex, times the number of images in which that vertex is visible, times the confidence of that vertex. This function is summed over all vertices in the mesh, and minimized using a coordinate descent method with a bounded finite difference Newton line search. The size of the finite difference increment is reduced between successive iterations by a factor of two until a minimum value is reached. After every step, a check is made to ensure that the objective function strictly decreases. The optimization was initialized with the output of an iterative closest point (ICP) alignment between the reference mesh and one of the submitted reconstructions. It was found that the result of the optimization was invariant to which sample reconstruction was selected for the ICP alignment. The quality of these alignments was validated by manually inspecting the reprojection of the full images; maximum reprojection errors were found to be on the order of 1 pixel, and usually substantially less.

4. Evaluation methodology

We now describe how we evaluate reconstructions by geometric comparison to the ground truth model.

Let us denote the ground truth model as G and the submitted reconstruction result to be evaluated as R . The goal of our evaluation is to assess both the *accuracy* of R (how close R is to G), and the *completeness* of R (how much of G is modeled by R). For the purposes of this paper, we assume that R is itself a triangle mesh.

To measure the accuracy of a reconstruction, we compute the distance between the points in R and the nearest points on G . Since R is a surface, in theory, we should construct measures that entail integration over R although in practice we simply sample R at its vertices.

A problem arises where G is incomplete. In this case, for a given point on R in an area where G is incomplete, the nearest point on G could be on its boundary or possibly on a distant part of the mesh. Rather than try to detect and remove such errors we instead compute nearest distances to G' , a hole-filled version of G , and discount points in R whose nearest points on G' are closest to the hole-filled regions. Figure 3(b) illustrates this approach. While this solution is itself imperfect, if the hole fills are reasonably

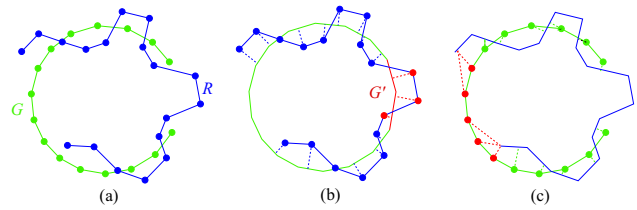


Figure 3. Evaluation of reconstruction R relative to ground truth model G . (a) R and G are represented as meshes, each shown here to be incomplete at different parts of the surface. (b) To compute accuracy, for each vertex on R , we find the nearest point on G . We augment G with a hole filled region (solid red) to give a mesh G' . Vertices (shown in red) that project to the hole filled region are not used in the accuracy metric. (c) To measure completeness, for each vertex on G , we find the nearest points on R (where the dotted lines terminate on R). Vertices (shown in red) that map to the boundary of R or are beyond an “inlier distance” from R to G are treated as not covered by R .

“tight,” this approach will avoid penalizing accurate points in R at the cost of discarding some possibly less accurate points that happen to match to the hole fill. In practice, we use the hole-filled surfaces generated by space carving [62] during surface reconstruction from range scans, and we perform many scans (approximately 200 per object), so that these hole fills are fairly close to the actual surface and constitute a small portion of the surface of the model. In addition, the mesh G has per-vertex confidence values indicating how well it was sampled by the scanner [62]; we ignore points on R that map to low confidence regions of G .

After determining the nearest valid points on G from R , we compute the distances between them. We compute the *signed* distances to get a sense of whether a reconstruction tends to under- or over-estimate the true shape. We set the sign of each distance equal to the sign of the dot product between the outward facing normal at the nearest point on G and the vector from that point to the query point on R .

Given the sampling of signed distances from the vertices of R to G (less the distances for points that project to hole fills of G'), we can now visualize their distribution and compute summary statistics useful in comparing the accuracy of the reconstruction algorithms. One useful example of such a statistic is to compute the distance d such that $X\%$ of the points on R are within distance d of G . When $X = 50$ for instance, this gives median distance from R to G . One such statistic is presented in Section 5.

To measure completeness, we compute the distances from G to R , i.e., the opposite of what we do for measuring accuracy. Intuitively, points on G that have no suitable nearest points on R will be considered “not covered”. Again, while we could measure the covered area by integration, we instead sample using the vertices of G , which are fairly uniformly distributed over G for our models. Unfor-

tunately, we cannot use the same idea for rejecting nearest points that we use for the accuracy metric, since, generally, a hole-filled R' is not available.

Instead, we propose an alternative completeness measure that simply reports the fraction of points of G that are within an allowable distance d of R ¹. The parameter d should be chosen to be large enough to accommodate “reasonable” errors in the reconstructions. A consequence of this measure is that unusually noisy reconstructions will tend to have lower completeness scores. Figure 3(c) illustrates the principle of the completeness measure.

5. Results

In this section, we present the results of our quantitative evaluation of six multi-view stereo reconstruction algorithms on the *temple* and *dino* datasets shown in Figure 1.

First, we briefly describe each algorithm. In an effort to cover the current state of the art, we sought to include new, recently published algorithms rather than evaluating classic methods from a few years ago. In addition to the six reported here, three other groups tried out the data but were not able to produce reasonable results and are therefore not included in the study.

Furukawa *et al.* [48] use wide-baseline stereo matching to recover the 3D coordinates of salient feature points, then shrink a visual hull model so that the recovered points lie on its surface, then refine the result using energy minimization. Goesele *et al.* [63] compute a depth map from each camera viewpoint (similar to [31]) and merge the results using *VRIP* [62]. Hernandez and Schmitt [31] first compute a depth map from each camera viewpoint and merge the results into a cost volume. They then iteratively deform a mesh, initialized at the visual hull, to find a minimum cost surface in this volume, also incorporating terms to fit silhouettes. Kolmogorov and Zabih [35] compute a set of depth maps using multi-baseline stereo with graph cuts, then merge the results into a voxel volume by computing the intersections of the occluded volumes from each viewpoint. Pons, Keriven, and Faugeras [26] compute a minimum cost surface by evolving a surface in a level-set framework, using a prediction-error measure. Vogiatzis, Torr, and Cipolla [19] compute a correlation cost volume in the neighborhood of the visual hull. A minimum-cost surface is then computed using volumetric min-cut.

We found that the different multi-view stereo reconstructions have sub-millimeter translational offsets with respect to each other. Relative to the accuracies of the best models, these offsets are quite significant. We postulate that these shifts are due in part to small errors in calibration, as shifts

¹Points on G that map to the boundary of R do not have a well-defined signed distance. We therefore treat half of those points as positive, and half as negative.

in the gantry can cause small offsets at different latitudes, but also reflect intrinsic differences between the algorithms. To compensate for these shifts, we first aligned the ground truth mesh (GTM) to each reconstruction using ICP, before computing the accuracy and completeness measures.

Table 1 summarizes the results of running our accuracy and completeness metrics on the aligned models that these six participants submitted. We used an accuracy threshold of 90%, i.e., an accuracy of $1.0mm$ means that 90% of the points are within one mm of the GTM. For completeness, we used an inlier threshold of $1.25mm$, i.e., a completeness of 95% means that 95% of the points are within $1.25mm$ of the GTM. We found that the accuracy and completeness rankings among the algorithms were relatively stable (see our web page [2] for results with other thresholds).

The accuracy of many of these methods is remarkable. Most methods consistently get sub-millimeter accuracy with very few outliers—and this is from images captured only at video resolution. Hernandez had the best accuracy on the temple datasets, with 90% of its points being within $0.36mm$ of the GTM on the full temple set. However, Hernandez consistently had one of the largest translational offsets among the algorithms (e.g., a shift of $0.6mm$ on the temple set)—if we had not normalized for such offsets the results would have changed significantly.

We were surprised how well methods did on the dino set, given that the only texture was due to subtle shading variations on the surface. Visual inspection of the reconstructions does show that even the best multi-view stereo results are noisier than the laser scanned GTM, indicating that there is potentially still room for improvement.

While accuracy numbers decreased with fewer images on the temple datasets, the dino results surprisingly show the opposite trend, with most methods doing better on the Ring than on the Full dino set. Due to the lack of texture on the dino, regularization likely plays a more central role.

Since most of the algorithms in this survey generate complete object models, the completeness numbers were not very discriminative. The primary exception is Goesele, whose reconstructions contain holes in low-confidence regions, and cause the lower completeness numbers for templeSparseRing (due to sparse view sampling) and the dino sets (due to areas with low texture).

Almost all of these algorithms exploited the fact that reasonable silhouettes were easily available via background thresholding on these data sets (Hernandez, Vogiatzis, and Furukawa require silhouettes to operate). An exception is Pons, which does not use silhouettes. Also, Goesele used silhouettes on the temple but not the dino datasets. We found these latter results encouraging, since silhouettes are not always available (e.g., the other datasets in Table 1).

We also note that the run-times of these algorithms varied dramatically, with Pons consistently the fastest (31 min-

	Temple			Dino		
	Full (317)	Ring (47)	SparseR. (16)	Full (363)	Ring (48)	SparseR. (16)
Furukawa [48]	0.65, 98.7%	0.58, 98.5%	0.82, 94.3%	0.52, 99.2%	0.42, 98.8%	0.58, 96.9%
Goesele [63]	0.42, 98.0%	0.61, 86.2%	0.87, 56.6%	0.56, 80.0%	0.46, 57.8%	0.56, 26.0%
Hernandez [31]	0.36, 99.7%	0.52, 99.5%	0.75, 95.3%	0.49, 99.6%	0.45, 97.9%	0.60, 98.5%
Kolmogorov [35]		1.86, 90.4%			2.80, 85.7%	
Pons [26]		0.60, 99.5%	0.90, 95.4%		0.55, 99.0%	0.71, 97.7%
Vogiatzis [19]	1.07, 90.7%	0.76, 96.2%	2.77, 79.4%	0.42, 99.0%	0.49, 96.7%	1.18, 90.8%

Table 1. Accuracy and Completeness results. The first number $0.xx$ measures *accuracy*: the distance d (in mm) such that 90% of the reconstruction is within d of the ground truth mesh (GTM). The second number $xx.x\%$ specifies *completeness*: the percent of points on the GTM that are within $1.25mm$ of the reconstruction. The number of views in each dataset is shown in parentheses in the table header.

utes on templeRing) and Goesele by far the slowest (more than a day on templeRing).

Our web page [2] contains many other statistics on these experiments, including run-times, unsigned and signed histograms of distances from reconstruction to ground truth model (and vice versa), cumulative histograms of distances, RMS error measures, and alignment offsets between the models and the GTM. While we lack space to show views of the reconstructions here, we strongly encourage readers to look at these renderings on our web pages; we feel that the accuracy numbers in Table 1 match quite well to the visual quality of the reconstructions.

6. Conclusions

This paper presented a taxonomy of multi-view stereo algorithms, new multi-view datasets registered with laser-scanned surface models, an evaluation methodology that measures accuracy and completeness, and a quantitative evaluation of some of the best-performing algorithms.

We are now preparing more challenging datasets with specularities, no silhouettes, etc., that we hope will help further advance the state of the art. We also plan to capture data at higher resolution and are investigating techniques such as industrial CT scanning to obtain higher accuracy ground truth. Finally, we are now opening the evaluation to allow other researchers to benchmark their algorithms against the best of breed techniques.

References

- [1] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1):7–42, 2002.
- [2] S. Seitz et al. Multi-view stereo evaluation web page. <http://vision.middlebury.edu/mview/>.
- [3] C. Dyer. Volumetric scene reconstruction from multiple views. In L. S. Davis, editor, *Foundations of Image Understanding*, pp. 469–489. Kluwer, 2001.
- [4] G. Slabaugh, B. Culbertson, T. Malzbender, and R. Shafer. A survey of methods for volumetric scene reconstruction from photographs. In *Intl. WS on Volume Graphics*, 2001.
- [5] T. Fromherz and M. Bichsel. Shape from multiple cues: Integrating local brightness information. In *ICYCS*, 1995.
- [6] S. Roy and I. Cox. A maximum-flow formulation of the N-camera stereo correspondence problem. In *ICCV*, pp. 492–499, 1998.
- [7] R. Szeliski and P. Golland. Stereo matching with transparency and matting. *IJCV*, 32(1):45–61, 1999.
- [8] S. Seitz and C. Dyer. Photorealistic scene reconstruction by voxel coloring. *IJCV*, 35(2):151–173, 1999.
- [9] P. Eisert, E. Steinbach, and B. Girod. Multi-hypothesis, volumetric reconstruction of 3-D objects from multiple calibrated camera views. In *ICASSP 99*, pp. 3509–3512, 1999.
- [10] J. De Bonet and P. Viola. Poxels: Probabilistic voxelized volume reconstruction. In *ICCV*, pp. 418–425, 1999.
- [11] K. Kutulakos and S. Seitz. A theory of shape by space carving. *IJCV*, 38(3):199–218, 2000.
- [12] K. Kutulakos. Approximate N-view stereo. In *ECCV*, vol. I, pp. 67–83, 2000.
- [13] A. Broadhurst, T. Drummond, and R. Cipolla. A probabilistic framework for the space carving algorithm. In *ICCV*, pp. 388–393, 2001.
- [14] R. Bhotika, D. Fleet, and K. Kutulakos. A probabilistic theory of occupancy and emptiness. In *ECCV*, vol. 3, pp. 112–132, 2002.
- [15] R. Yang, M. Pollefeys, and G. Welch. Dealing with textureless regions and specular highlights – a progressive space carving scheme using a novel photo-consistency measure. In *ICCV*, pp. 576–584, 2003.
- [16] T. Bonfort and P. Sturm. Voxel carving for specular surfaces. In *ICCV*, pp. 591–596, 2003.
- [17] A. Treuille, A. Hertzmann, and S. Seitz. Example-based stereo with general BRDFs. In *ECCV*, vol. II, pp. 457–469, 2004.
- [18] G. Slabaugh, B. Culbertson, T. Malzbender, and M. Stevens. Methods for volumetric reconstruction of visual scenes. *IJCV*, 57(3):179–199, 2004.
- [19] G. Vogiatzis, P. Torr, and R. Cipolla. Multi-view stereo via volumetric graph-cuts. In *CVPR*, pp. 391–398, 2005.
- [20] O. Faugeras and R. Keriven. Variational principles, surface evolution, PDE’s, level set methods and the stereo problem. *IEEE Trans. on Image Processing*, 7(3):336–344, 1998.

- [21] J.-P. Pons, R. Keriven, O. Faugeras, and G. Hermosillo. Variational stereovision and 3D scene flow estimation with statistical similarity measures. In *ICCV*, pp. 597–602, 2003.
- [22] S. Soatto, A. Yezzi, and H. Jin. Tales of shape and radiance in multiview stereo. In *ICCV*, pp. 974–981, 2003.
- [23] H. Jin, S. Soatto, and A. Yezzi. Multi-view stereo beyond lambert. In *CVPR*, vol. 1, pp. 171–178, 2003.
- [24] Y. Duan, L. Yang, H. Qin, and D. Samaras. Shape reconstruction from 3D and 2D data using PDE-based deformable surfaces. In *ECCV*, vol. 3, pp. 238–251, 2004.
- [25] H. Jin, S. Soatto, and A. Yezzi. Multi-view stereo reconstruction of dense shape and complex appearance. *IJCV*, 63(3):175–189, 2005.
- [26] J.-P. Pons, R. Keriven, and O. Faugeras. Modelling dynamic scenes by registering multi-view image sequences. In *CVPR*, vol. II, pp. 822–827, 2005.
- [27] P. Fua and Y. Leclerc. Object-centered surface reconstruction: Combining multi-image stereo and shading. *IJCV*, 16:35–56, 1995.
- [28] A. Rockwood and J. Winget. Three-dimensional object reconstruction from two-dimensional images. *Computer-Aided Design*, 29(4):279–285, 1997.
- [29] L. Zhang and S. Seitz. Image-based multiresolution shape recovery by surface deformation. In *SPIE: Videometrics and Optical Methods for 3D Shape Measurement*, pp. 51–61, 2001.
- [30] J. Isidoro and S. Sclaroff. Stochastic refinement of the visual hull to satisfy photometric and silhouette consistency constraints. In *ICCV*, pp. 1335–1342, 2003.
- [31] C. Hernandez and F. Schmitt. Silhouette and stereo fusion for 3D object modeling. *CVIU*, 96(3):367–392, 2004.
- [32] T. Yu, N. Xu, and N. Ahuja. Shape and view independent reflectance map from multiple views. In *ECCV*, pp. 602–616, 2004.
- [33] R. Szeliski. A multi-view approach to motion and stereo. In *CVPR*, vol. 1, pp. 157–163, 1999.
- [34] S.-B. Kang, R. Szeliski, and J. Chai. Handling occlusions in dense multi-view stereo. In *CVPR*, vol. I, pp. 103–110, 2001.
- [35] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *ECCV*, vol. III, pp. 82–96, 2002.
- [36] C. Zitnick, S.-B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. *ACM Trans. on Graphics*, 23(3):600–608, 2004.
- [37] P. Gargallo and P. Sturm. Bayesian 3D modeling from images using multiple depth maps. In *CVPR*, vol. II, pp. 885–891, 2005.
- [38] M.-A. Drouin, M. Trudeau, and S. Roy. Geo-consistency for wide multi-camera stereo. In *CVPR*, vol. I, pp. 351–358, 2005.
- [39] G. Vogiatzis, P. Torr, S. M. Seitz, and R. Cipolla. Reconstructing relief surfaces. In *BMVC*, pp. 117–126, 2004.
- [40] G. Zeng, S. Paris, L. Quan, and F. Sillion. Progressive surface reconstruction from images using a local prior. In *ICCV*, pp. 1230–1237, 2005.
- [41] R. Szeliski. Prediction error as a quality metric for motion and stereo. In *ICCV*, pp. 781–788, 1999.
- [42] S. Savarese, H. Rushmeier, F. Bernardini, and P. Perona. Shadow carving. In *ICCV*, pp. 190–197, 2001.
- [43] M. Okutomi and T. Kanade. A multiple-baseline stereo. *TPAMI*, 15(4):353–363, 1993.
- [44] A. Prock and C. Dyer. Towards real-time voxel coloring. In *Image Understanding WS*, pp. 315–321, 1998.
- [45] P. Narayanan, P. Rander, and T. Kanade. Constructing virtual worlds using dense stereo. In *ICCV*, pp. 3–10, 1998.
- [46] A. Laurentini. The visual hull concept for silhouette-based image understanding. *TPAMI*, 16(2):150–162, 1994.
- [47] S. Sinha and M. Pollefeys. Multi-view reconstruction using photo-consistency and exact silhouette constraints: A maximum-flow formulation. In *ICCV*, pp. 349–356, 2005.
- [48] Y. Furukawa and J. Ponce. High-fidelity image-based modeling. Technical Report 2006-02, UIUC, 2006.
- [49] C. Stewart. Robust parameter estimation in computer vision. *SIAM Reviews*, 41(3):513–537, 1999.
- [50] S. Baker, T. Sim, and T. Kanade. When is the shape of a scene unique given its light-field: A fundamental theorem of 3D vision? *TPAMI*, 25(1):100–109, 2003.
- [51] T. Tasdizen and R. Whitaker. Higher-order nonlinear priors for surface reconstruction. *TPAMI*, 26(7):878–891, 2004.
- [52] J. Diebel, S. Thrun, and M. Bruenig. A Bayesian method for probable surface reconstruction and decimation. *ACM Trans. on Graphics*, 25(1), 2006.
- [53] H. Saito and T. Kanade. Shape reconstruction in projective grid space from large number of images. In *CVPR*, vol. 2, pp. 49–54, 1999.
- [54] G. Slabaugh, T. Malzbender, B. Culbertson, and R. Schafer. Improved voxel coloring via volumetric optimization. TR 3, Center for Signal and Image Processing, 2000.
- [55] O. Faugeras, E. Bras-Mehlman, and J.-D. Boissonnat. Representing stereo data with the Delaunay triangulation. *Artificial Intelligence*, 44(1–2):41–87, 1990.
- [56] A. Manassis, A. Hilton, P. Palmer, P. McLauchlan, and X. Shen. Reconstruction of scene models from sparse 3D structure. In *CVPR*, vol. 1, pp. 666–673, 2000.
- [57] D. Morris and T. Kanade. Image-consistent surface triangulation. In *CVPR*, vol. 1, pp. 332–338, 2000.
- [58] C. J. Taylor. Surface reconstruction from feature based stereo. In *ICCV*, pp. 184–190, 2003.
- [59] D. Wood, D. Azuma, K. Aldinger, B. Curless, T. Duchamp, D. Salesin, and W. Stuetzle. Surface light fields for 3D photography. In *SIGGRAPH*, pp. 287–296, 1996.
- [60] W.-C. Chen, J.-Y. Bouguet, M. Chu, and R. Grzeszczuk. Light field mapping: Efficient representation and hardware rendering of surface light fields. *ACM Trans. Graphics*, 21(3):447–456, 2002.
- [61] J.-Y. Bouguet. Camera calibration toolbox for Matlab. http://www.vision.caltech.edu/bouguetj/calib_doc/.
- [62] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH*, pp. 303–312, 1996.
- [63] M. Goesele, B. Curless, and S. Seitz. Multi-view stereo revisited. In *CVPR*, 2006. To appear.