Region-based monocular reconstruction: does it miss the points?

S. Cord Melton, Lilia Markham, and Zachary Dodds

Abstract — **In this paper we investigate a spectrum of approaches to region-based 3d visual reconstruction. On one hand, space-carving techniques require no prior environmental assumptions; on the other, triangulation based approaches offer built-in regularization to mitigate noise. Even at a coarse level, these region-based approaches are a promising complement to the successes of monocular mapping via point and line features. What is more, region-based approaches offer flexibility in scenes with dynamic texture (foliage) or without texture (walls), as well as on low-cost or educational platforms which offer only presegmented images.**

I. Motivation and Context

Vision offers perhaps the highest bandwidth-to-cost ratio to robotic systems. In theory at least, one camera is a complete sensor suite. This potential, countered by pixels' close coupling of lighting, scene, and optics, has spurred decades of reconstructive research - reaching as far back as Shakey and the Stanford Cart. The progress in 3d reasoning from image streams has accelerated over the past five years: today's algorithms untangle even monocular data into consistent and accurate environmental representations [1-4]. Results such as [5] are, quite frankly, inspiring.

Yet these remarkable systems exhibit a relatively narrow focus in their design: they rely on sparse, accurately reconstructed image features. Stepping back, one might consider a spectrum of reconstructive approaches in which features' precision trades off against their density, as sketched in Figure 1.

Work to date emphasizes the bottom-up paradigm: SIFT, SURF, KLT, et al. provide locally distinct feature points, triangulate them into 3d via robust statistics, and hang textures onto the result [6,7]. An alternative path runs up-then-across in Figure 1: segmentation algorithms [8,9] produce regions whose interframe correspondences yield a rough set of 3d surfaces. From those surfaces, texture mapping and further image processing sculpt more accurate world representations -- but only as required or desired.

Indeed, with this work we argue that despite the successes of environmental assembly, environmental sculpting offers a complementary approach useful in situations where precise feature matching can fail:

(1) with densely or dynamically textured natural surfaces - such as bark, foliage, or water
(2) with featureless surfaces typical of some indoor office environments
(3) for tasks like obstacle avoidance, where the density of the world's representation supersedes its absolute accuracy
(4) when only segmented regions are available, not the images themselves.

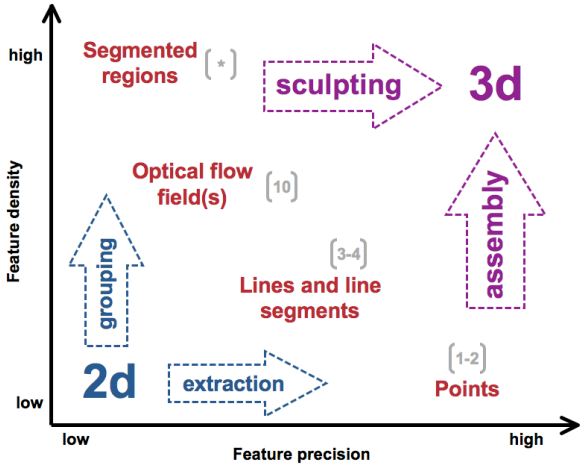

Figure 1. A taxonomy of visual reconstruction approaches. Bibliographic citation numbers place algorithms according to the precision and density of the underlying 2d features they use and 3d primitives they create. Whereas most algorithms first extract accurately localized features, usually points, and then assemble them into a 3d cloud, this work [*] transforms image regions into coarser estimates of world structure.

This fourth use case has become increasingly common as low-cost platforms for research and education proliferate. For example, the Mach 5 [11], KIPR's XBC [12], and Lego's NXT [13] offer access only to presegmented data through Cognachrome [14], CMUcam [15], or Mindsensors [16] interfaces. Other widely accessible platforms offer "blob" tracking as an option for a pedagogically accessible introduction to vision [17,18]. We contend that blobs, whether optional or required, do not preclude 3d reasoning on such platforms.

A. Context and Contributions

Figure 1's counterpoint of world-assembly and world-sculpting is nothing more than the computer vision community's dichotomy between structural reconstruction and space carving. This work simply leverages the fact that

Manuscript received, September 1, 2007. This work was supported in part by the U.S. NSF DUE grants (0451293 and 0536173) as part of the 2007 summer CS REU at Harvey Mudd College.

S. Cord Melton is a junior at the University of Chicago. Chicago, IL 60637 USA (email: author@ boulder.nist.gov).

L. Markham is a senior at Harvey Mudd College, Claremont, CA 91711 USA. (email: lmarkham@hmc.edu).

Z. Dodds is an associate professor of Computer Science, Harvey Mudd College, Claremont, CA 91711 USA (phone: 909-607-1813; fax: 909-607-8364; email: dodds@cs.hmc.edu).

an incomplete and/or coarse space-carved approximation can offer a great deal of utility to a robotic agent. Perhaps because of its computational cost and pose-accuracy requirements, pure space carving has not seen substantial use for robotic environmental modeling. For instance, [19] eases the burden with human-segmented and -matched image patches; [20] space-carves automatically, but through a one-dimensional camera retina.

In contrast to many reconstructive algorithms which use direct range sensing along with vision [21,22], we follow [1] by using only monocular data and approximate odometry, such as that available on the low-cost platforms cited above. Our work concentrates on cameras whose optical axis remains parallel to the ground plane; thus, it allows fewer camera freedoms than [1]. This ground-plane restriction is not inherent; rather it reflects a natural starting point for this work. Thus, as a first step, this paper sets a foundation for useful region-based environmental reasoning via

(1) A spectrum of algorithmic variations for coarse reconstruction via segmented image sequences (Sec. 2)
(2) Performance metrics and results for the tasks of obstacle avoidance and landmark reaquisition (Sec. 3)
(3) Data sets and source code openly available from [23].

In the end, the techniques presented across Figure 1's axes are neither mutually exclusive nor comprehensive. Though this work is in its early stages, we feel its algorithms show that -- depending on visual and environmental conditions -- roboticists' toolkits will find use for both assembling and sculpting approaches to monocular resonstruction. Ultimately, hybrids will combine the advantages of each.

II. Region-based Reconstruction

A. *Inputs: 2d image regions*

We begin by segmenting all input images using Edison [24]. Edison only segments; our system proceeds to compute region correspondences across the image series. Regions are matched between consecutive images based on location, color, size, and shape characteristics. At this 2d-reasoning stage, "objects" are simply collections of corresponding image regions, (hopefully) representing the same physical object viewed from different poses, as shown in Figure 2.

Figure 2. ***(L to R)*** Two images (out of 75), with their segmentations.

B. *(Alg.1) Cylindrical reconstruction via triangulation*

We consider three algorithms for creating a coarse 3d representation from these 2d "objects." The first algorithm augments traditional triangulation [25] with the assumption that the centers of each object's 2d regions correspond to a consistent 3d point. Certainly, this is not true; it implicitly presumes that objects are spherical and unoccluded. Our approach estimates the radius of this presumed sphere from the camera's calibration and the 2d regions' widths. Because of the ground-plane assumption, we represent these objects as cylinders, not spheres.

C. *(Alg.2) Fencepost reconstruction*

A partial refinement of the cylindrical triangulation of Alg. 1 instead builds a collection of planes by triangulating the left and right endpoints of each pair of corresponding 2d regions. This "fencepost" approximation builds a sheaf of vertical planar patches in 3d whose convex hull, in the absence of noise, contains the object of interest. This approximation well approximates convex obstacles with polygonal cross-section, e.g., many man-made structures.

D. *(Alg.3) Space-carving via fenceposts*

The third variation further leverages these *fenceposts*, i.e., the left and right vertical boundaries of objects from a particular viewpoint. Here, the world is presumed solid; obstacle volumes are carved away outside of those fenceposts' projections, as illustrated in Figure 3's top-down view.

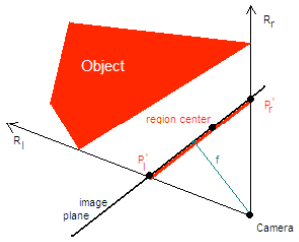

Figure 3. Rays R_l and R_r project through an object's "fenceposts" for a particular image at P_l and P_r.

Step-by-step details of these algorithms appear in [26]; here, we augment these brief descriptions with the reconstructions in Figure 4 and section III. The 3d modeling, in part, uses the powerful computational geomtery library CGAL [27] and OpenGL for rendering the results.

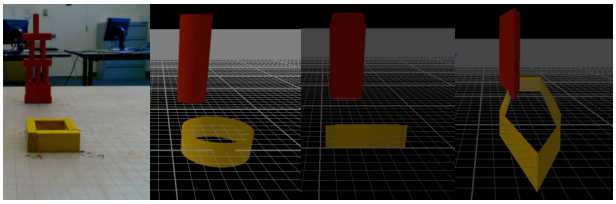

Figure 4. ***(L to R)*** An input image and example reconstructions of the two objects using cylindrical, plane-sheaf, and space-carved models.

III. Results and Evaluation

A. *Metrics for assessing reconstructions*

The promise of 3d representations is its support of both low-level tasks, e.g., obstacle avoidance, as well as higher-level ones, e.g., navigation and loop closure. We evaluate our algorithms with metrics motivated by these goals.

To assess our 3d maps' support for navigation-based tasks, we define *recklessness* as the fraction of actual obstacle area not enclosed by the corresponding reconstructed object.

Paranoia is the fraction of reconstructed object area not enclosed by the actual obstacle. Both metrics are error measurements, ranging from 0.0 (perfect) and 1.0 (no intersection). Figure 5 provides two numeric examples of each metric, based on only two images viewing the scene in Figure 4 almost head-on.

To assess our maps' ability to help reaquire landmarks, e.g., for loop-closure or global path planning, we use the following algorithm:

(1) choose a novel viewpoint; move the camera to it in both the real and rendered worlds

(2) segment both the real and rendered images taken from this viewpoint and compute region correspondences

(3) we define the *reacquisition error* to be the distance

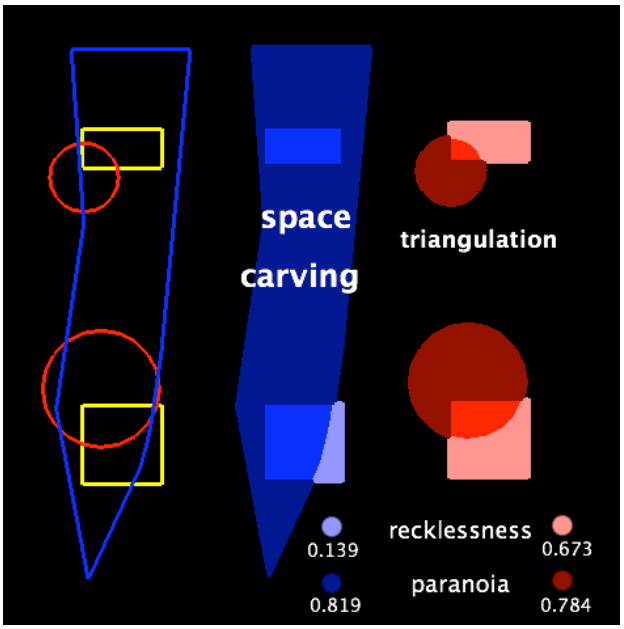

Figure 5. ***(L to R)*** Top-down projections of space-carved (blue) and cylindrical (red) reconstructions based on two input images. Real object footprints are in yellow. The *recklessness* and *paranoia* values track the tradeoff between aggressive and conservative occupancy estimates.

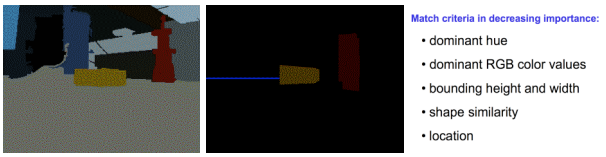

Figure 6. ***(L to R)*** Real and model-rendered segmented images from a novel camera viewpoint, along with the statistics used to quantify the differences between the resulting 2d regions; their sum constitutes the *reacquisition error* metric, stated for each object at bottom right.

between the real and rendered region parameters (Fig. 6)

B. Results

The blocks-world testbed of Figures 2-6 has provided a convenient starting point to measure and compare these region-based approaches to 3d reconstruction: it provides ground truth and factors out the the efficacy of the "off-the-shelf" segmentation system we employ.

Trading off paranoia and recklessness

Figure 7's data illustrate how cylindrical models outperform space-carving in both of these navigation-based metrics for large numbers of images, though for small numbers, triangulation is clearly superior in paranoia and space carving is superior in recklessness.

Triangulation benefits from additional viewpoints because they increase the accuracy of the centroid estimate. The radius of the reconstructed cylinder tends to remain relatively constant as the number of images increase. Space carving, in contrast, performed less well as the number of viewpoints increased. Given ideal segmentation, space carving would continue to improve in paranoia while never sacrificing perfect recklessness.

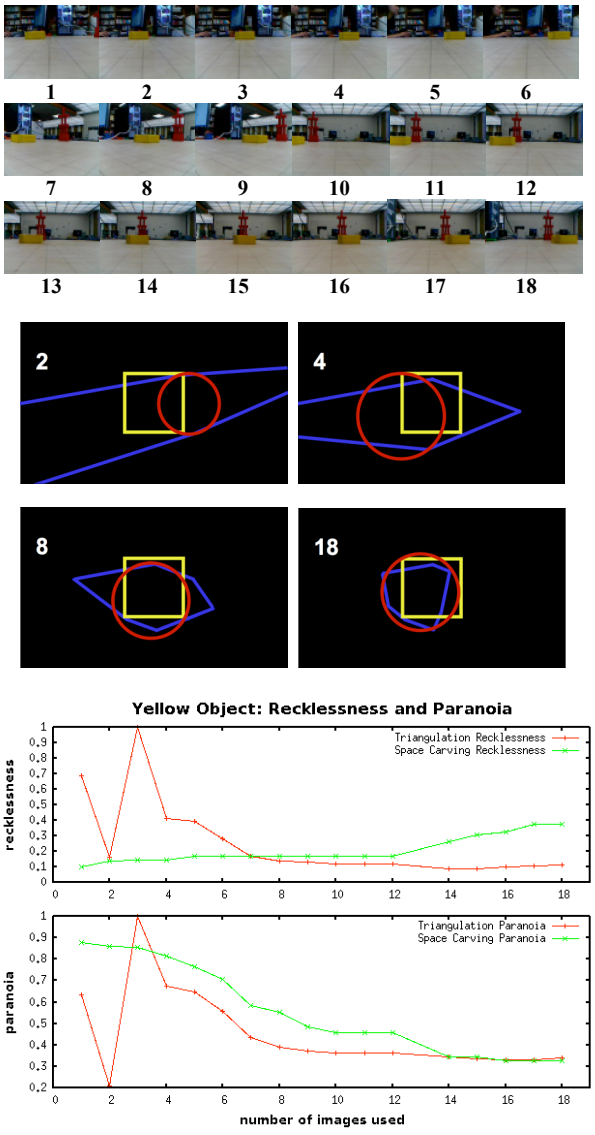

Figure 7. ***(top)*** A portion of the input image sequence ***(middle)*** top-down views of reconstruction results at image numbers 2, 4, 8, and 18 for space-carving (blue) and cylindrical triangulation (red) of the yellow object, whose reference position is also shown ***(bottom)*** plotting *recklessness* and *paranoia* values across this sequence.

Correct segmentation would guarantee that any space inside the original object would be seen by the camera as such and would never be carved away. However, as we have imperfect segmentation, sometimes pieces of the true segment are lost. This means that space inside the actual object is carved away. As the number of images increases, inevitably so does the number of incorrect segments, and so does the recklessness of the space-carving reconstruction.

Measuring landmark reacquitision

Reacquiring landmarks is crucial to augmenting robots' reactive capabilities with higher-level reasoning about larger-scale environments. The ability to "close the loop" upon reentering a previously-viewed area forms the basis for both SLAM and path-planning algorithms. We have measured the *reacquisition error* of our reconstruction algorithms by selecting viewpoints at which to compare real and rendered images of the scene. Each of the Section II's three algorithms contributed a rendered image, based on its reconstruction from a full 75-image sequence. Figure 7's 18-image subsequence shows part of that data: from there, the camera continued to circle the scene. Figure 8 provides additional snapshots and examples of reacquisition error.

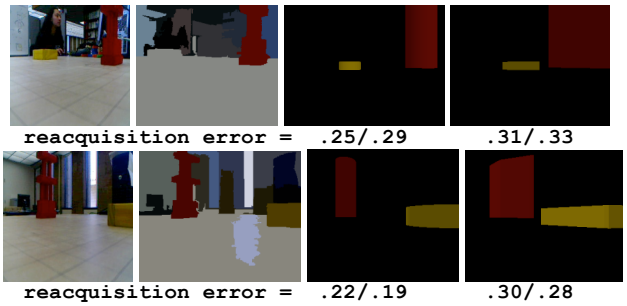

Figure 8. ***(L to R)*** Two additional images of the blocks-world scene, their segmentations, cylindrical and space-carved reconstructions, projected onto the reference viewpoint. Below each landmark are its values for landmark-reacquisition error, computed using the criteria in Figure 6 and then normalized to a value between 0 (no error) and 1. As in these two examples, the triangulation-based reconstruction yields better landmark recognition overall than space-carved models (Figure 10).

A taxonomy of landmark-matching errors

Figure 8 also suggests a best-score matching algorithm: one that pairs a rendered landmark with the real image region that minimizes its reacquistion error. Because the *actual* object that gives rise to a rendered landmark may or may not be present in the real image, this leads to six possibilities for this matching. Three of them are ***correct***:

- a match seen in both images
- no match, as the rendering is incorrect
- no match in either the rendered or real image

By the same token, three possibilities are ***incorrect***:

- a mismatched landmark
- a missed match, though present in both images
- a missed match, as the rendering is incorrect

Figure 9 summarizes these six possible results for landmark-matching. For each image in the original sequence, we rendered the scene from that camera's pose and then ran the matching algorithm for landmark-reacquisition. The results appear in Figure 10.

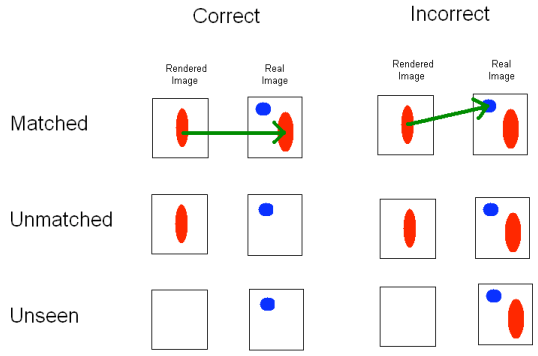

Figure 9. The three correct (left) and three incorrect (right) possibilities when matching 2d region-landmarks from rendered and real images.

REACQUISITION via		CYLINDERS			PLANE-SHEAVES			SPACE CARVING		
Landmark	Result Type	Inc.	Tot.	Acc.	Inc.	Tot.	Acc.	Inc.	Tot.	Acc.
Red	Matched	0	60	100.00%	2	60	96.67%	4	59	93.22%
	Unmatched	0	4	100.00%	0	5	100.00%	2	8	75.00%
	Unseen	0	9	100.00%	0	8	100.00%	0	6	100.00%
	All	0	73	**100.00%**	2	73	**97.26%**	6	73	**91.78%**
Yellow	Matched	1	37	97.30%	1	38	97.37%	0	36	100.00%
	Unmatched	1	24	95.83%	0	23	100.00%	1	27	96.30%
	Unseen	0	12	100.00%	0	12	100.00%	0	10	100.00%
	All	2	73	**97.26%**	1	73	**98.63%**	1	73	**98.63%**
Average				**98.63%**			**97.95%**			**95.21%**

Figure 10. Summary of landmark-reacquisition data for the red and yellow objects. These excellent results underscore how sensitive this metric is to the particulars of the environment. Yet even for these very distinctive landmarks, errors did occur, as shown in Figure 11.

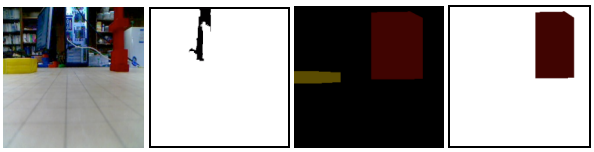

Figure 11. An example of a landmark mismatch even within our blocks-world dataset. The rendered red space-carved landmark matched most closely (0.34) with a shadowed background region. The low confidence difference of 0.003 (compare Figure 12) suggests that the system should place almost no faith in this match.

Figure 10's excellent results reflect as much about the simplicity of the scene as they do about the algorithm itself. Even so, errors did occur – and with increasing visual complexity, those errors would become even more common. Figure 11 shows one such error, in which the large difference in the size between the space-carved model (right) and the actual segmented image (left) leads to an incorrect correspondence.

Although further tuning could fix Figure 11's particular mismatch, trying to avoid all such situations is fruitless. Rather, it is more important for a spatial-reasoning system to

maintain an estimate of its confidence in a particular landmark-reacquisition match. We quantify this *confidence* as the difference between the best and second-best match-scores for each rendered landmark, as seen in Figure 12.

Not all images are equal: *thoroughness* and *novelty*

Background conditions can change dramatically as the camera moves about a scene. Indeed, Figure 12's data illustrate how sensitive our confidence measure is to camera pose. A key problem with that plot is its equal treatment of each image – in practice, the system relies on some images far more than others in building its 3d reconstructions. For instance, a single view from one side of a landmark will contribute almost as much as a large collection of closely-spaced images from the other side. In order to distinguish these different contributions from different input images, our system defines *novelty* and *thoroughness* for its inputs.

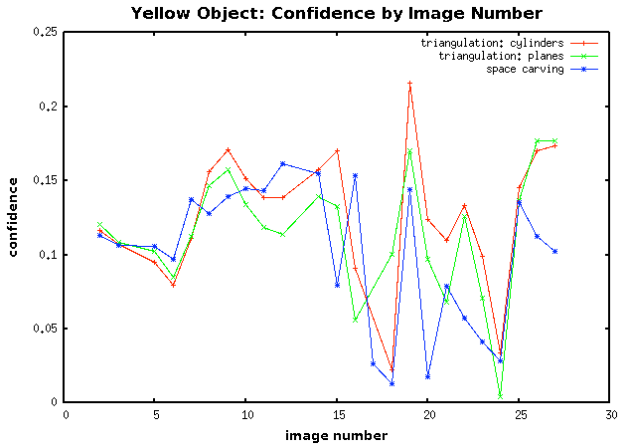

Figure 12. Plot of landmark-reacquisition confidence (the best minus the second-best match score) in sequential images, for each of the three reconstruction algorithms. These huge variations prompted explicit consideraton of each image's contribution to the reconstructed 3d object models. The *thoroughness* and *novelty* metrics were the result (Fig. 12b)

Consider an image set s and an object o with perimeter p, and the portion s' of p visible in at least one image in s. Then, the *thoroughness* of s with respect to o is s' / p, that is, the fraction of the perimeter of o visible in at least one image in s . To compute novelty, consider additionally an image n, the portion n' of p visible in n, and the portion sn' of p visible in both n and at least one image in s. The *novelty* of n is $(n'-sn')/n'$, that is, the fraction of n' that is not visible in at least one image in s.

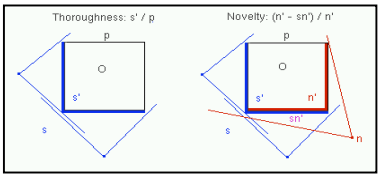

Figure 12b. Ilustrating *thoroughness* and *novelty* metrics for images that contribute to a 3d scene reconstruction.

Novelty and thoroughness, in turn, provide a backdrop against which to evaluate our three reconstruction algorithms. Figure 13 shows the confidence of the yellow object's reacquisition plotted against the novelty of the novel viewpoint. It confirms the natural trend for confidence to decrease as novelty increases for all of the algorithms, with space-carving the least confident of the three. Figure 14's plot of the yellow landmark model's recklessness and paranoia reprises the data from Figure 7. Plotted against the thoroughness of the image set used, rather than image index, this new plot replaces Figure 7's anomalies with the smooth improvements expected with additional data.

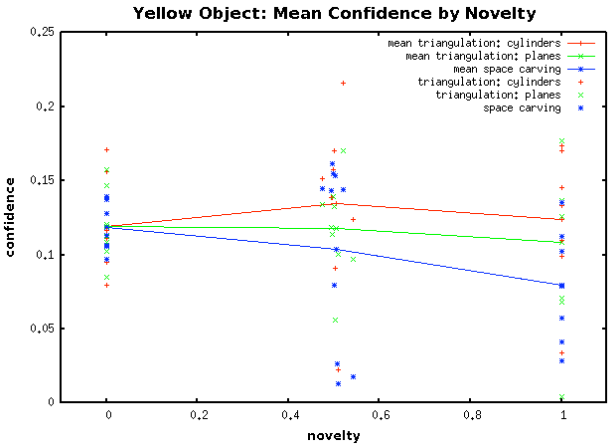

Figure 13. Plot of the data from Figure 12, organized by the novelty of the contributing image. As expected, individual background differences yield a wide spread for each distinct novelty value. Yet in five of the six segments the expected downward trend is apparent.

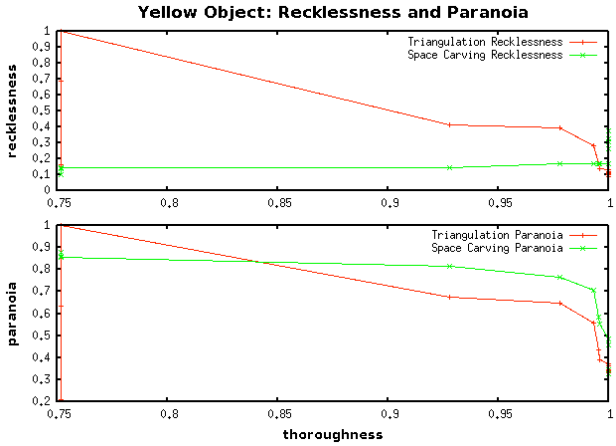

Figure 14. Plot of the data from Figure 7, organized according to the thoroughness of the underlying image set. Here, paranoia decreases as more and more of the object is seen: the carving algorithm excises more of its generous initial assumption and the additional data tightens the triangulation algorithm's estimates by lessening the effects of noise. Reckless decreases similarly for the cylindrical models, but because space-carving can never recapture a region that has been carved away, recklessness there can only increase, albeit slowly.

In the end, the coarsely reconstructed cylindrical and plane-sheaf landmarks perform better than space-carved landmarks – for obstacle avoidance (*recklessness*) and path planning (*paranoia*). Landmark-reacquisition further bears this out, both in overall match scores and in those matches' *confidence*. These differences persist even when accounting for the *thoroughness* and *novelty* of the underlying images.

We do not feel these results, even as unambigous as they are, impugn the importance or potential of space-carving as a resource for region-based reconstruction. Rather, the message seems to reinforce the remarkable power of good prior models. After all, the two landmarks on which we focused are quite close to cylinders – and even closer to the convex hull of a sheaf of planes. Space carving's lack of prior assumptions enable it to handle for more complex landmarks without special accommodation.

On the other hand, space carving can not recover well from noise or errors in the 2d segmented regions that are the input to the algorithm. There is no mechanism for "reclaiming" space that has already been carved away. This suggests a hybrid approach, in which coarse triangulated estimates are both refined by space carving and a fall back for ensuring that space carving does not carve away too much. By leveraging this experience with evaluating reconstructive approaches, we hope to help articulate and assess such hybrid algorithms in the future.

IV. Perspective

The blocks-world examples presented here underscore how early in development we are with this work. Even at this preliminary stage, however, the approaches, metrics, and results suggest that sparse-feature-based reconstruction is not the only means to reason visually about the 3d world. Indeed, it would be surprising if the single strategy of interpolating from a small subset of accurately-localized image data sufficed for all visual conditions. The dual approach -- coarsely placing and/or sculpting objects from the visible environment and refining as needed -- opens up complementary capabilities.

A lingering open problem is one of representation. What data structures will best mediate these differing sources of information and enable refinement of the 3d hypotheses they generate? It seems that multiple *confidence resolutions* will need to interleave with varying spatial resolutions to capture varying image contributions.

We look forward to the robust algorithms that will emerge from combining these approaches into hybrid reasoning that smoothly incorporates both visually distinct features and more diffusely delimited data. Success in human-robot interaction will ultimately depend on robots' ability to reason about the 3d world in which humans perceive themselves. We will strive to help robotic systems realize this ability through the inexpensive, powerful, and richly varied channel of monocular vision.

References

[1] A. J. Davison, I. Reid, N. Molton and O. Stasse. "MonoSLAM: Real-Time Single Camera SLAM" *IEEE Trans. PAMI* (accepted for publication), 2007.

[2] D. L. Cardon, W. S. Fife, J. K. Archibald, and D. J. Lee. "Fast 3D reconstruction for small autonomous robots" Proc. *IECON*, Nov 2005.

[3] P. Smith, I. Reid, and A. J. Davison. "Real-Time Monocular SLAM with Straight Lines" *Proceedings, British Machine Vision Conference*, 2006.

[4] Y.-Q. Cheng, X.G. Wang, R.T. Collins, E.M. Riseman, and A.R. Hanson. "Three-Dimensional Reconstruction of Points and Lines with Unknown Correspondence across Images" International Journal of Computer Vision 45(2) pp. 129-156, 2001.

[5] P. Mordohai, J.-M. Frahm, A. Akbarzadeh, B. Clipp, C. Engels, D. Gallup, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, H. Towles, G. Welch, R. Yang, M. Pollefeys and D. Nistér, "Real-Time Video-Based Reconstruction of Urban Environments", *Proceedings of 3DARCH: 3D Virtual Reconstruction and Visualization of Complex Architectures*, Zurich, Switzerland, July, 2007

[6] D. Nistér. "Automatic dense reconstruction from uncalibrated video sequences" PhD Thesis, Royal Institute of Technology KTH, Stockholm, Sweden, ISBN 91-7283-053-0, March 2001.

[7] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, R. Koch, "Visual modeling with a hand-held camera" *International Journal of Computer Vision* 59(3), 207-232, 2004.

[8] D. Comanicu and P. Meer: "Mean shift: A robust approach toward feature space analysis." *PAMI*, 24, 603-619, May 2002.

[9] P. Felzenszwalb and D. Huttenlocher. "Efficient Graph-Based Image Segmentation" *IJCV*, 59(2) September 2004.

[10] Weber, J and Malik, J. "Rigid Body Segmentation and Shape Description from Dense Optical Flow under weak perspective" PAMI, 139-143, Feb 1997.

[11] Newton Labs' Mach 5 platform (accessed 7/15/07) www.newtonlabs.com/mach5.htm

[12] R. LeGrand, K. Machulis, D. P. Miller, R. Sargent, and A. Wright. "The XBC: a modern low-cost mobile robot controller" Edmonton, Canada, IROS 2005.

[13] M. McNally, F. Klassner, and C. Continanza. "Exploiting MindStorms NXT: Mapping and Localization Projects for the AI Course" FLAIRS-20 Key West, FL May 7-9, 2007.

[14] R. Sargent, B. Bailey, C. Witty, and A. Wright. "Dynamic object capture using fast vision tracking" *AI Magazine*, Spring 1997, Volume 18, No. 1

[15] A. Rowe, C. Rosenberg, and I. Nourbakhsh. "A low-cost embedded color vision system" IROS 2002, vol1. pp. 208-213.

[16] Mindsensors NXT camera, mindsensors.com

[17] J. Bruce, T. Balch, and M. Veloso. "Fast and inexpensive color segmentation for interactive robots" IROS 2000 volume 3, pp. 2061 – 2066.

[18] B. Gerkey, R. T. Vaughan and A. Howard. "The Player/Stage Project: Tools for Multi-Robot and Distributed Sensor Systems". ICAR 2003, Coimbra, Portugal, pp. 317-323

[19] R. Ziegler, W. Matusik, H. Pfister, and L. McMillan. "3D reconstruction using labeled image regions" 2003 Eurographics Symp. on Geometry Processing, pp.1-12, 2003.

[20] A. Eppendahl and A. Ojamaa, "Seeing Empty Space in an Environment without Silhouettes" *Proceedings, AMiRE* 2005.

[21] D. Hähnel, W. Burgard, and S. Thrun. "Learning compact 3D models of indoor and outdoor environments with a mobile robot" *Robotics and Autonomous Systems*, 44:15-17, 2003.

[22] Ohno, K.; Tadokoro, S. "Dense 3D map building based on LRF data and color image fusion" *Proceedings, IEEE Int. Conf. On Intelligent Robots and Systems* (IROS), pp. 2792-2797, Aug. 2005.

[23] www.cs.hmc.edu/twiki/bin/view/Robotics/ThreeDFromTexture

[24] Edge Detection and Image Segmentation (EDISON) System, Robust Image Understanding Lab, Center for Advanced Information Processing, Rutgers University, 2002.

[25] R. Hartley and P. Sturm. "Triangulation" *Computer Vision and Image Understanding* 68(2), pp. 146-157, 1997.

[26] L. Markham, S. C. Melton, and Z Dodds. "Robot control via region-based 3d reconstruction" In *Proceedings, Intelligent Systems and Control '07* Cambridge, MA; Nov. 2007.

[27] **CGAL**, www.cgal.org : Computational Geometry Algorithms Library