

Dense 3D Mapping with Monocular Vision

Kamil Wnuk, Faith Dang, and Zachary Dodds
Computer Science Department,
Harvey Mudd College, Claremont, CA, USA
dodds@cs.hmc.edu

Abstract

This paper describes a recently developed toolkit for creating environmental maps of indoor environments from image sequences. First, the chosen map representation – texture-mapped planes located in a single coordinate frame – is placed within the broader context of appearance-based and sparse feature-based visual mapping approaches. Drawing from the well-developed field of structure-from-motion, an algorithm is presented for building such maps with a limited sensor suite: monocular vision and odometry. Maps validating this approach are presented, along with task-independent metrics made possible by the representation. We evaluate our maps according to these metrics and conclude with implications for future sensor-limited spatial reasoning.

Keywords: vision-based robot mapping, monocular vision, map representation, map evaluation, structure from motion, software toolkits for robot vision

1 Introduction and Related Work

Despite cameras' long history as a robotic sensor, e.g. [1,2], important questions in visual map-building still remain. A fundamental issue is that of representation: when should spatial reasoning occur directly via image features and when should it occur in a Euclidean coordinate system attached to the robot or its surroundings? This paper describes ongoing work in building representations that support both types of reasoning, i.e., dense, metrical visual maps.

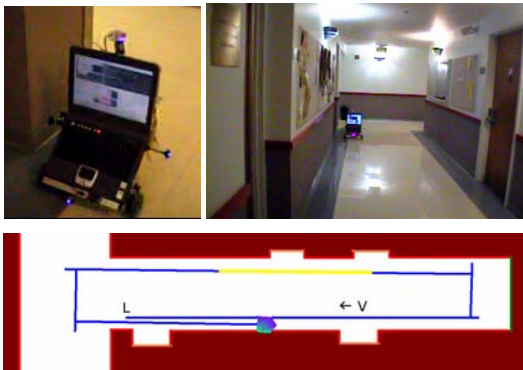


Figure 1: A laptop-controlled ER1. In each image the robot is at **L**; **V** is the viewpoint shown at right. The USB camera and odometry contribute to map-building, and short-range IRs help with exploration.

Our focus is on map-building with monocular vision and odometry; figure 1 depicts a typical setting. Such a limited sensor suite offers several advantages: it is inexpensive, at ~US\$300 off-the-shelf, there is no multiview calibration to maintain, it facilitates miniaturization [3], and it allows a variety of task-specific form factors, e.g. [4,5].

Research into visual mapping with these types of platforms can be categorized by choices in map representation. Figure 2 locates several comparable vision-based robotic systems along two axes: the *density* of maps' stored image information and the *modeling tradeoff* between images and objects. Maps to the upper right of this taxonomy preserve a sparse set of 2d image regions and their summary statistics, e.g., SIFT features [6,7], SSD patches [8], and image histograms [7,9], all of which benefit from fast matching against similarly processed input images. Global geometric relationships in these systems are ad hoc, typically via environment-specific adjacency graphs [10].

Figure 2's upper-left systems similarly represent 3d space primarily in terms of images, not objects, but they employ subspace methods such as PCA [11,12] to distill image data without token extraction. In order to explain visual changes without modelling the camera, extended environments need considerable visual exploration [13,14,15], but can handle substantial image variation as a result [16]. A hybrid image-based approach interpolates among appearance variations in patches of high local edge density to identify 3d poses from 2d image regions [17].

Maps at the lower right of the taxonomy maintain global 3d coordinates of a sparse set of visually distinct features, e.g., points [18], lines [19,20], or color transitions [21]. Though naturally supported by stereo [18,22,23], monocular approaches also exist [24]. Both can benefit from the anticipatory rendering available from the computed 3d structure, e.g., [25].

Indeed, environmental rendering -- either for human or algorithmic use -- is an important motivation for methods that explicitly map the positions *and* visual properties of scene surfaces, i.e., the lower left corner of figure 2. Dense 3d maps further allow explicit reasoning about the environment's geometry, e.g., for motion planning or map evaluation via ground-truth comparisons. Perhaps most important is such maps' ability to integrate disparate sensory information by not tying environmental representation to a particular sensor. Recent systems have exploited this strength to build impressive models with stereo cameras [23], laser range finders [26,27], panoramic vision [28,29] or a combination of these [30,31,32].

This paper focuses on the more modest capability of monocular vision, a sensor suite less fully explored for robotic mapping. In part, this stems from its well-established niche within vision's structure-from-motion (SFM) subfield. Indeed, in this work we adapt traditional SFM approaches [33,34,35] to datasets more typical of robotic applications, i.e., odometrically annotated scenes of planar indoor environments. Other researchers have pointed single cameras toward the floor [36] and ceiling [37]; our approach is similar, but exploits the visual and geometric salience of the walls between those two extremes.

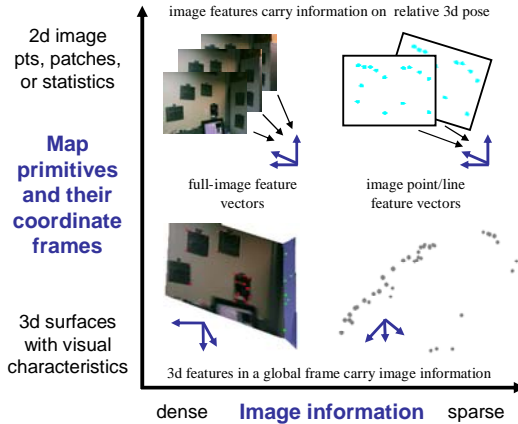


Figure 2: A taxonomy of visual map-building approaches organized by coordinate systems used and the density of visual information maintained.

2 Monocular map-building

Here we describe our mapping toolkit, noting its differences from traditional structure-from-motion approaches. Figure 3 charts the flow of information through the system; section 3 will present intermediate and final results validating the approach. The data sets and complete source code (C++ for Windows) including visualization tools are available at www.cs.hmc.edu/wart.

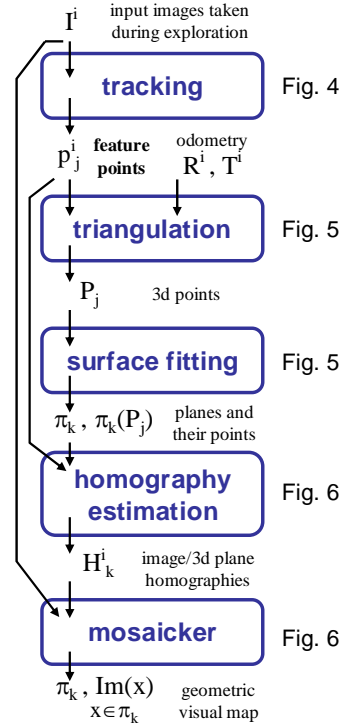


Figure 3: A summary of the steps our system takes to build dense visual maps from monocular data and odometry, adapted from the SFM approach in [33]. Indicated figures depict intermediate results.

Processing begins with feature identification and tracking; we employ Birchfeld's KLT tracker [38] on the input image sequence, I^i . Images are taken ~ 1 second apart during autonomous exploration, corresponding to about 20cm of camera displacement. The tracker's outputs, p_j^i , are the (x,y)-coordinates of the j th feature point in the i th image, I^i . It is important that these indices apply across the full image sequence, so that if feature j does not appear in I^i , the corresponding p_j^i is marked as null. Though this differs from the assumptions made for single-object modelling by standard techniques such as [39], bookkeeping for feature dropouts is crucial when modelling an extended environment within which occlusions necessarily occur.

These image features p_j^i are next mapped into a global frame of reference by triangulation. Odometry provides extrinsic camera rotations R^i and translations T^i for each image. Offline calibration [40,41] yields the camera's intrinsic parameters, K , which are presumed constant. Neither odometry nor K are strictly necessary, however, since all of the parameters may be estimated up to a scale factor from image information alone [33]. In either case, least-squares estimation allows all available image data to contribute to the 3d coordinates, P_j , of the image

features p_j^i in the odometry's global coordinate frame:

$$p_j^i = K [R^i | -R^iT] P_j \quad (1)$$

As equation 1 suggests, correspondence is preserved between image and 3d point locations. The output P_j , in turn, are fed to a plane extractor, where the system assumes the environment is piecewise planar. For indoor, hallway-dominated scenes such as those in figure 1, this is a reasonable assumption. Note that a dense stereo matcher, e.g., one of the many evaluated in [42], would eliminate this restriction.

The plane extractor hypothesizes clusters of planes with RANSAC [43] and judges each cluster with several heuristics: points' 3d proximity, the tendency of indoor environments' surfaces to be parallel or perpendicular to one another, and the low likelihood that a point is shared by more than one plane. These heuristics contribute to a fitness function that scores assignments from points to planes: $\pi_k(P_j) = 1$, if P_j is on plane π_k , and $\pi_k(P_j) = 0$ otherwise. An additional cluster π_\emptyset gathers outliers. Of 5,000 possible assignments of the point set to planes π_k , the best fit is preserved as the planar environmental map.

While the P_j themselves suffice to represent environmental structure, the advantage of this plane-fitting step is that a much denser, but still geometrically accurate, visual representation is possible through image mosaicking. Homographies H_k^i are computed via least-squares between images I^i and planes π_k that share feature points [33]:

$$\pi_k(P_j) = H_k^i p_j^i \quad (2)$$

Here $\pi_k(P_j)$ represents the local coordinates of P_j within a 2d coordinate frame attached to π_k . Assuming Lambertian surfaces, the homographies provide radiosity correspondence between each plane's image I_k and the input images, I^i :

$$I_k(x) = I^i(H_k^{i-1}(x)) \quad (3)$$

This mapping of color to the estimated 3d structure is the job of the mosaicker. Our implementation creates a single composite image for each plane π_k by averaging corresponding in-bounds pixels from each input image. The result is an image I_k that holds the map's surface properties.

3 Results and Evaluation

These texture-mapped planes constitute a dense, 3d representation of the environment. Figure 7 shows a partial map of a lab resulting from a short run at a distance of 2m. Intermediate figures present the output from each of figure 3's modules. The robot's path, the left-facing camera poses, and two representative snapshots with tracked features are shown in figure 4. Figure 5 depicts two views of the 3d point clouds triangulated from these tracked features, along with similar views of the best-fit planes. Points on the foreground computer monitor have been removed as outliers.

Figure 6 illustrates a texture for the "red" plane (the whiteboard) created by homography estimation using the points in green, followed by mosaicking. This texture is then rendered in figure 7's final map. Fully saturated green pixels represent areas for which no image information is available. These regions could be used by the system, e.g., to guide further exploration, though such behavior has not been explored in our implementation. In contrast to approaches elsewhere in figure 2's taxonomy, the texture-mapped representation illustrated here allows the rendering of viewpoints far from those observed during exploration, e.g., a top-down view.

As noted in the introduction, a further benefit of these maps is the ability to evaluate them against ground truth. This kind of task-independent evaluation, we feel, is an important and underinvestigated facet of reusable robotics algorithms. Figure 9 summarizes the feature position errors between hand-measured "truth" and the structure estimated in figure 7's map.

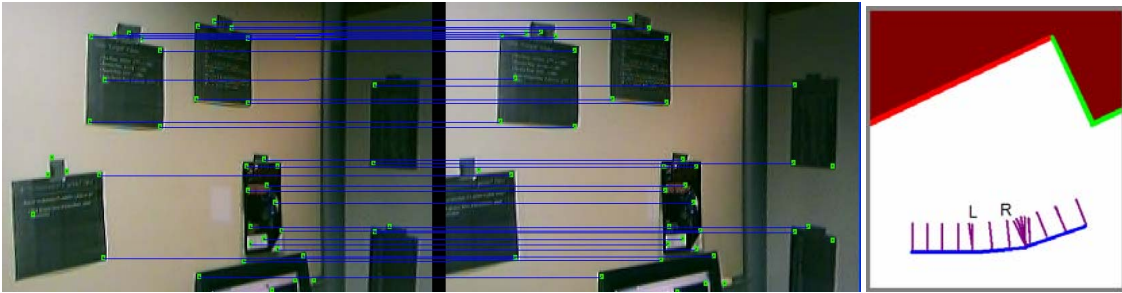


Figure 4: Corresponding KLT features found by the tracking module between two of the lab sequence's input images. The two views (L and R) are marked within the robot's complete 1.5m run in the schematic on the right

A second, mapping run appears in Figure 8, where the robot has explored a hallway (the yellow path portion shown in figure 1) by wall-following and obstacle avoidance with short-range IR sensors. Although figure 8's map consists of only one plane, it demonstrates the ability of the mosaicker to create single-image representations of environments too large for the camera to capture at one.

The error discrepancies demonstrate another benefit of the maps of figures 7 and 9: an explicit measure of available visual resolution can be computed for each rendered point. By accounting for the 3d location of the imaged surface and the pose of the camera that gave rise to its estimated radiosity, this resolution can be expressed for each contributing image in units of pixels/cm. With an available resolution of ~ 2.1 pixels/cm along the whiteboard in the lab corner map, figure 9 indicates that about 7 pixels of discrepancy between model and truth have accumulated through the map-making process (and in estimating by hand the 3d points the tracker found in the images). Further reasoning based on such maps can take advantage of this measure of available visual resolution, e.g., to motivate uncertainty models for probabilistic visual/spatial reasoning.

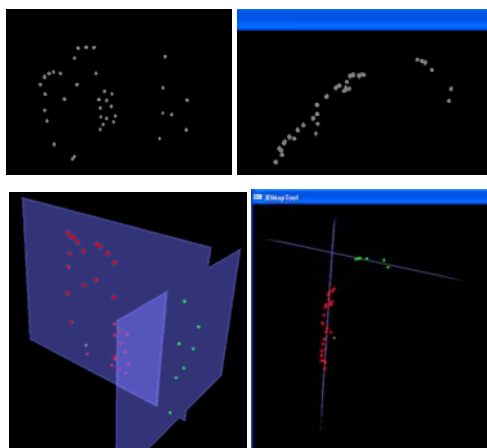


Figure 5: Two views of the point cloud resulting from triangulation (top) and plane assignments (bottom). The colors are as in figure 4's floorplan.

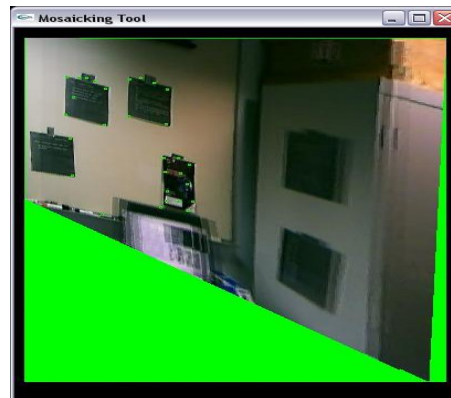


Figure 6: An uncropped, mosaicked texture from the lab sequence using only feature points from the whiteboard plane. Alignment is quite consistent on that plane, but not on other planes (as expected).

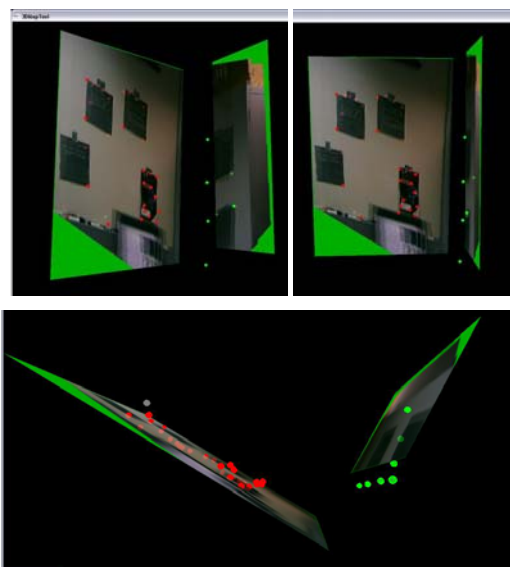


Figure 7: Renderings of the final "lab-corner" map with cropped textures from the two dominant planes. A benefit of the combined image-based and/geometric representation is the possibility of rendering views far from any of the input images, e.g., the lower image.



Figure 8: Rendering of the "hallway" map. Though only a single plane, this map fuses the 10 images from the red portion of the path in figure 1. To the left appear the visual repercussions of a rotation not modeled by the robot's odometry. Such artefacts argue for adjusting the odometric poses based on future data, i.e., SLAM.

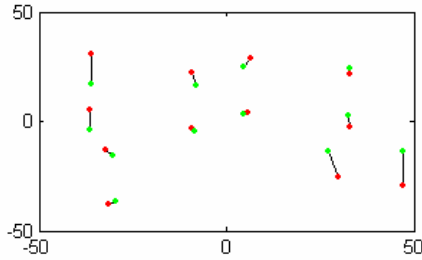


Figure 9: The average feature location error taken over 12 features from the whiteboard plane is 2% of the distance to the surface imaged, or 3.5 cm. Real locations are green with corresponding estimates red.

4 Perspective

Section 3 highlights some of the advantages of the mapping system described in this work: task-independent evaluation, anticipatory rendering from unseen poses, and a complete yet compact representation of both geometric and visual data. On the other hand, our system has several limitations: the planarity assumption, the lack of past pose adjustment given future image data (SLAM), and computational cost – the system now runs offline at about ~10s per image. Though they use different representations and/or sensor suites, the successes of [24,27,31] demonstrate that each of these problems stem from our current implementation and are not inherent in the proposed approach.

Ultimately a multiresolution representation, as in [26], will likely emerge to combine sparse features for initial indexing and dense data for more involved visual/spatial reasoning and display. Both object- and image-based representations will play a role in such maps, as will axes not represented in figure 2. For instance, Sim suggests a “spectrum of prior information” [44], which would generalize the planar assumptions made here. Efforts that fuse structure-from-motion and autonomous mapping, we believe, will continue to advance the capabilities of low-cost robotic systems.

5 Acknowledgements

Support for this work comes from National Science Foundation award NSF CCLI A&I #0411176 and funds provided by Harvey Mudd College.

6 References

- [1] Fikes, R., Hart, P., and Nilsson, N., “Learning and executing generalized robot plans,” *Artificial Intelligence*, 3 (4), pp 251-288 (1972).
- [2] Moravec, H., “Towards automatic visual obstacle avoidance,” *Proc., 5th IJCAI*, vol. vision-1, p. 584 (1977).
- [3] Grabowski, R., Navarro-Serment, L. E., and Khosla, P. K., “Small is beautiful: an army of

- small robots,” *Scientific American*, 289 (5), pp 63-67+ (Nov 2003).
- [4] Jackson, B., Burt, I., Kratochvil, E., and Papanikolopoulos, N., “A control system for teams of off-the-shelf scouts and megascouts,” *Proc. 11th IEEE Mediterranean Conf. on Control and Automation*, (June 2003).
- [5] Rybski, P. E., Roumeliotis, S. I., Gini, M., Papanikolopoulos, N., “Building topological maps with sensor-limited miniature mobile robots,” *Proc., IROS*, pp 194-199 (Oct 2003).
- [6] Lowe, D. G., “Distinctive image features from scale-invariant keypoints,” *Int. Journal of Computer Vision*, 60 (2), pp 91-110 (2004).
- [7] Kosecka, J. and Li, F., “Vision based topological markov localization,” *Proc., ICRA*, pp. 1481-1486 (Apr 2004).
- [8] Hager, G. D., Kriegman, D., Yeh, E., and Rasmussen, C., “Image-based prediction of Landmark Features for Mobile Robot Navigation,” *Proc., ICRA*, pp 1040-1046 (Apr 1997).
- [9] Ulrich, I. and Nourbakhsh, I., “Appearance-based place recognition for topological localization,” *Proc., ICRA*, pp 1023-1029 (Apr 2000).
- [10] Kosecka, J. and Yang, X., “Global localization and relative positioning based on scale-invariant features,” *Proc., ICPR*, to appear (Aug 2004).
- [11] Nayar, S. K., Nene, S. A., and Murase, H., “Subspace methods for robot vision,” *IEEE Tr. Rob. and Automation*, 12 (5), pp 750-758 (1996).
- [12] Nayar, S. K., Murase, H., and Nene, S. A., “Learning, positioning, and tracking visual appearance,” *Proc., ICRA 1994*, pp 3237-3246 (May 1994).
- [13] Pourraz, F. and Crowley, J. L. “Continuity properties of the appearance manifold for mobile robot position estimation,” *Proc., 2nd IEEE Workshop on Perception for Mobile Agents*, Ft. Collins (June 1999).
- [14] Winters, N. and Santos-Victor, J., “Information sampling for vision-based robot navigation,” *Journal of Robotics and Autonomous Systems*, 41 (2-3), pp 145-159 (Nov 2002).
- [15] Krose, B. J. A., Vlassis, N., Bunschoten, R., and Motomura, Y., “A probabilistic model for appearance-based robot localization,” *Image and Vision Computing* 19 (6), pp 381-391 (Apr 2001).
- [16] Bischof, H., Jogan, M., Leonardis, A., and Wildenauer, H., “Mobile robot localization under varying illumination,” *Proc., 16th ICPR*, R. Kasturi, D. Laurendeau, and C. Suen, Eds., vol II, pp 741-745, IEEE (2002).
- [17] Sim, R. and Dudek, G., “Learning generative models of scene features,” *Int. Journal of Computer Vision*, 60 (1), pp 45-61, (2004).

- [18] Saez, J. M., Penalver, A., Escolano, F., "Compact mapping in plane-parallel environments using stereo vision," *Progress in Pattern Recognition, Speech, and Image Analysis*, A. Sanfeliu and J. Ruiz-Shulcloper, Eds., LNCS 2905, Springer, pp 659-666 (2003).
- [19] Burschka, D., Eberst, C., Robl, C., and Farber, G., "Vision-based exploration of indoor environments," *Robust Vision for Vision-Based Control of Motion*, M. Vincze and G. D. Hager, Eds., IEEE Press, pp77-95, (2000).
- [20] Neira, J., Ribeiro, M. I., and Tardós, J. D., "Mobile robot localization and map building using monocular vision," *Proc., 5th Int. Symp. on Intelligent Robotic Systems*, Stockholm, Sweden, pp 275-284 (July 1997).
- [21] Yuen, D. C. K., and MacDonald, B. A., "Natural landmark based localisation system using panomaic images," *Proc., ICRA*, pp. 915-920 (May 2002).
- [22] Se, S., Lowe, D., and Little, J., "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *Int J. Robotics Research*, 21 (8), pp 735-758 (Aug 2002).
- [23] Iocchi, L., Konolige, K., and Bajracharya, M., "Visually Realistic Mapping of a Planar Environment with Stereo," *Experimental Robotics VII LNCIS*, D. Rus and S. Singh, Eds., Springer, pp 521-532 (June 2003).
- [24] Davison, A. J., "Real-time simultaneous localisation and mapping with a single camera," *Proc. ICCV*, (Oct 2003).
- [25] Molton, N. D., Davison, A. J., and Reid, I. D., "Locally planar patch features for real-time structure from motion," *Proc., British Machine Vision Conf.*, to appear (Sep 2004).
- [26] Montemerlo, M. and Thrun, S., "A multi-resolution pyramid for outdoor robot terrain perception," *Proc. AAAI*, pp 464-469 (July 2004).
- [27] Nuchter, A., Surmann, H., Lingemann, K., Hertzberg, J., and Thrun, S., "6D SLAM with an application in autonomous mine mapping," *Proc. ICRA*, pp1998-2003 (Apr 2004).
- [28] Bunschoten, R. and Krose, B., "Robust scene reconstruction from an omnidirectional vision system," *IEEE Tr. on Robotics and Automation*, 19 (2), pp. 351-357 (Apr 2003).
- [29] Thompson, S., Matsui, T., and Zelinsky, A., "Localisation using automatically selected landmarks from panoramic images," *Proc., Australian Conference on Robotics and Automation*, Melbourne, (Sep 2000).
- [30] Cobzas, D., Zhang, H., and Jagersand, M., "A panoramic model for robot predictive display," *Proc., 16th Vision Interface (VI '03)*, (June 2003).
- [31] Biber, P., Andreasson, H., Duckett, T., Schilling, A., "3D modeling of indoor environments by a mobile robot with a laser scanner and panoramic camera," *Proc., IROS*, to appear (Oct 2004).
- [32] Pervolz, K., Nuchter, A., Surmann, H., and H., Hertzberg, J., "Automatic reconstruction of colored 3d models," *Proc., Robotik 2004*, VDI-Berichte 1841, Munich, pp 215-222 (June 2004).
- [33] Ma, Y., Soatto, S., Kosecka, J., and Sastry, S. *An Invitation to 3D Vision: From Images to Models* Springer Verlag, December 2003.
- [34] Debevec, P., Taylor, C., and Malik, J., "Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach," *Proc., SIGGRAPH*, pp. 11-20 (Aug 1996).
- [35] Pollefeys, M., Vergauwen, M., Verblest, F., Cornelis, K., Tops, J., and Koch, R., "Visual modelling with a handheld camera," *Int. J. of Computer Vision*, 59 (3), pp 207-232 (Oct 2004).
- [36] Hajjdiab, H. and Laganieri, R., "Vision-based multirobot simultaneous localization and mapping," *Proc., 1st Canadian Conf. on Computer and Robot Vision*, pp 155-162 (May, 2004).
- [37] Dellaert, F., Burgard, W., Fox, D., and Thrun, S., "Using the condensation algorithm for robust, vision-based mobile robot localization," *Proc. CVPR*, pp 594-600 (June 1999).
- [38] Shi, J. and Tomasi, C., "Good features to track," *Proc., CVPR*, pp 593-600 (June 1994). Code available 8.1.04 at www.ces.clemson.edu/~sjb/klf.
- [39] Morita, T. and Kanade, T., "A sequential factorization method for recovering shape and motion from image streams," *Tr. on Pattern Analysis and Machine Intelligence*, 19 (8), pp 858-867 (Aug 1997).
- [40] Zhang, Z., "Flexible camera calibration by vieweing a plane from unknown orientations," *Proc. ICCV*, pp. 666-671 (Sep 1999). Code available 8.1.04 from the open source CV library at www.intel.com/research/mrl/research/opencv/.
- [41] Heikkila, J and Silven, O., "A four-step camera calibration procedure with implicit image correction," *Proc. CVPR*, pp. 1106-1111 (June 1997). Code also in OpenCV, cited above.
- [42] Scharstein, D. and Szeliski, R., "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. Journal of Computer Vision*, 47 (1-3), pp 7-42 (Apr 2002).
- [43] Fischler, M. A. and Bolles, R. C., "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Comm. Of the ACM*, 24 (6), pp 381-395 (June 1981).
- [44] Sim, R. *On visual maps and their automatic construction* Ph.D. thesis, McGill University, p.14 (Jan 2004).