Object-based Visual SLAM: How Object Identity Informs Geometry

Antonio H. P. Selvatici and Anna H. R. Costa
Laboratório de Técnicas Inteligentes — LTI
Escola Politécnica, University of São Paulo
São Paulo, SP

Frank Dellaert
College of Computing
Georgia Institute of Technology
30332-0280 Atlanta, GA

Abstract

Objects are rich information sources about the environment. A 3D model of the objects, together with their semantic labels, can be used for camera localization as well as for cognitive reasoning about the environment. However, traditional frameworks for scene reconstruction usually map a cloud of points using structure-from-motion techniques, but do not provide objects representation. On the other side, robotics object-based mapping mainly focus on adding cognitive representations to a metric or topologic map built using traditional SLAM techniques. In this work we propose a framework for environment modeling by representing the objects in the scene, detected by an object recognition and segmentation technique. The key idea is to incorporate the resulting image segments and labels into a global inference engine in order to build simple geometric models for the objects. For now, we consider the perfect object recognition case, where we know the exact object identities, testing our approach using coarsely hand-annotated images captured by a robot carrying an omnidirectional camera. We found that the resultant object locations and sizes are fully compatible with what is expected, and the inferred robot trajectory is improved when compared to that recovered using odometry only.

1. Introduction

Mixed geometric and semantic 3D models of the environment are useful either for human visualization or as a map for autonomous visual systems that must localize themselves and reason about the surrounding environment. When these systems interact with human beings, like in Augmented Reality (AR) applications or service robots, their world representation should share symbols with that of humans. Furthermore, there are evidences that people themselves use objects to represent indoors spaces [1]. In this context, object-based 3D models of the environment, like the one depicted in Fig. 1, are very suitable for providing landmarks for camera localization as well as cognitive entities for reasoning, while containing important elements for place visualization systems to build on.

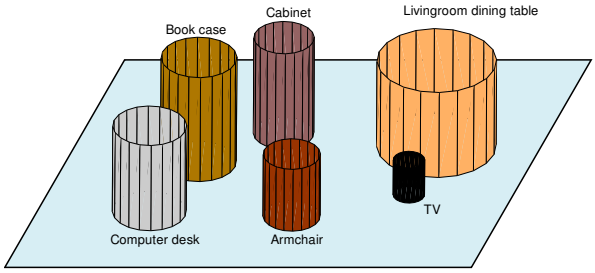


Figure 1. Example of a simple 3D model of the objects in a living room, showing their heights and average widths, as well as class labels.

In this work, we build simple object-based 3D models of the environment from an image sequence captured by a camera placed on a mobile robot, and also recover its trajectory. This is done by integrating the output of an object recognition and segmentation algorithm into a usual structure-from-motion (SFM) inference engine. In this sense, this work is different from traditional object-based semantic mapping approaches in robotics, since they tend to concentrate on the cognitive environment modeling problem [1][2]. In those works, the geometric aspect of the objects modeling is then simplified to informing positions in a certain reference frame, discarding the information about object sizes and other visual information, not used in the mapping process.

This work also differs from traditional visual SLAM approaches, which map only some interest points in the environment [3][4][5], without pro-

viding objects representation. Our goal is to build light-weight 3D models of objects in the environment, together with a semantic label indicating their class (e. g., clock, TV set, table, etc.) and identity, provided by object recognition. Slightly closer to our approach, [6] also incorporates some higher-level entities in the map, corresponding to recognized planar patches. However, these patches only indicate that certain points in the map belongs to a specific structure, making their associations throughout images more reliable due to this additional patch-level of matching test. Our work improves on that by using apparent size information rather than just image point positions.

The main idea we explore is that, if we roughly know the average real-world size of the objects belonging to a certain class, the apparent size of an instance in the image leads to a coarse range estimate from the camera to the object. Moreover, if we can also make assumptions about the object location, e. g., that a coach is more likely to be on the floor plane then onto a table, the detected object image also gives us clues about the camera pose. Figure 2 illustrates how it works. This idea was also used by Hoiem and others [7] to recover the camera viewpoint and the position of determined objects on the floor plane from single images, while performing object recognition using a third-party technique. However, here we relax the restrictions on the object locations and recover larger scale models using image sequences.

In this work, we focus on the scene modeling and camera pose estimation problem, which is solved by a traditional efficient inference method used in SFM. The problem is first modeled as a sparse linearized least-squares (LS) one, and then efficiently solved by means of QR factorization [8]. This method has been recently extended to the robot SLAM domain [9], [10], replacing many filtering approaches with great advantage. Although the general framework presented can assume different assumptions about the confidence on the object recognition technique used, we consider the perfect object recognition case, making use of an annotated image sequence database to test our approach.

The paper is organized as follows. In section 2 we present a general framework for our approach, from which we derive the specific model we use in this work. In section 3 we give details about the specific probabilistic models and inference algorithm we adopted. Experimental results are presented in section 4, while the conclusions of this work are discussed in section 5.

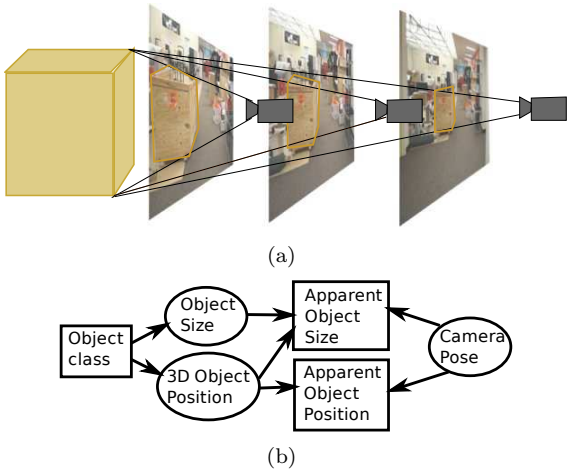


Figure 2. Illustration of how knowledge about the object size can be used to estimate the relative distance from the camera. a) The decreasing size of the wooden box in the images combined with prior knowledge about the size of wooden boxes informs about its increasing distance to the camera. b) Graphical probabilistic model illustrating the dependencies among object position and size, and camera pose (unknowns, represented by ellipses) given object recognition output (rectangles).

2. General Framework for Object-based SLAM

Our objective is to perform Maximum a Posteriori (MAP) inference to obtain a 3D model θ of the environment objects, using measurement data Z provided by object recognition in an image sequence. The MAP estimate θ^* is defined as

$$\theta^* = \arg\max_{\theta} P(\theta|Z) \tag{1}$$

We propose to take an approach where object classes, locations, and associated geometries are inferred together with the camera trajectory and orientations, tightly coupling these variables. In this case, $\theta = (X, M)$ [9], where X is the sequence of camera poses and M is the model, which includes, in addition to object locations L, their geometry G and their class labels C. Thus, let us define $M \triangleq \{o_j\}_{j=1}^{N}$, where each object $o_j \triangleq (l_j, g_j, c_j)$ is described by an object location l_j, the geometry g_j, and the class label c_j.

The data $Z = \{z_k\}_{k=1}^{K}$ provided by the object recognition system is assumed to comprise the apparent contour and position of the objects detected in the image sequence $\mathcal{I} = \{I_i\}_{i=0}^{T}$. Hence, we assume we always have

$$z_k = (u_k, s_k, \bar{c}_k) \tag{2}$$

where each object detection z_k provides a 2D location u_k, the respective apparent shape s_k, and the detected class $\bar{c}_k$. We can also define the correspondence variable, $J \triangleq \{(i_k, j_k)\}_{k=1}^{K}$, which is a mapping from measurement indices k to image indices i and object indices j, such that o_{j_k} is the object detected in image I_{i_k} giving raise to the measurement z_k. Depending on the set up, odometer readings about the camera movement, $V = \{v_i\}_{i=1}^{T}$, may also be available. This is the case in our experiments.

Depending on the confidence we have in the object recognition algorithm, there are several possible assumptions we can make regarding whether real object class labels C and correspondence J is known or not. The different choices are:

1) correspondence J known, class labels C known
2) correspondence J known, class labels C unknown
3) correspondence J unknown, class labels C known
4) correspondence J unknown, class labels C unknown

In the case correspondence is known, it is implied that we know the number of objects N. However, if correspondence is unknown, N itself becomes an object of inference.

If the object recognition technique used is reliable enough, we may assume that each object detected is uniquely identified, and also that its class label is recognized perfectly. This is the case where choice 1 applies, since object identities give raise to the correspondence J. In this work we investigate only the first situation, where both correspondence and class labels are known.

2.1. Known Correspondence and Class Labels

Assuming that we know the object classes and respective identities, as we stated above, we can adopt a similar approach used in traditional SFM, but now our structure include also object geometry. The posterior (1) can be expressed by:

$$P(\theta|Z;J) \propto P(Z,\theta|J) = P(X,M|J)P(Z|X,M;J) \quad (3)$$

where $P(X, M|J)$ is a prior density on trajectory and object models, which might include odometer information, if available. $P(Z|X, M; J)$ is the measurements likelihood.

At this point, we have to explicit the assumed variables and measurements relationships in order

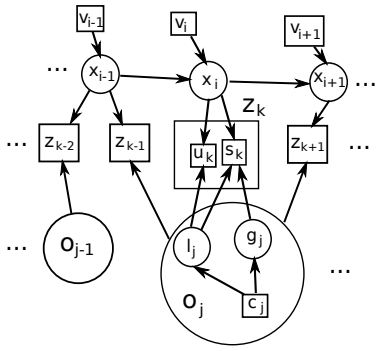


Figure 3. Fragment of the Bayesian Network that represents the probabilistic model for object-based SLAM when class labels and correspondence are known. The rectangles represent known variables, and the circles represent the unknowns. We basically assume that objects class give a rough idea about its geometry, and may give clues about its possible locations. We also assume that the apparent shape is independent of the object location in image given the unknowns. This model does not admit occlusion among objects, neither moving objects.

to define our model. These relationships are represented in Fig. 3. Basically, we assume that object classes influence their geometry, and possibly their locations. Since the model M comprises the set of object locations L, 3D geometry G, and the known objects class labels C, hence the prior density can be written as

$$\begin{aligned} P(X, M|J) &= P(X)P(L|C)P(G|C) \\ &= P(X) \prod_{j=1}^{N} \{P(l_j|c_j)P(g_j|c_j)\} \end{aligned} \quad (4)$$

If odometer information V is available, the prior on the camera poses is given by

$$P(X) = P(x_0) \prod_{i=1}^{T} P(x_i|x_{i-1}, v_i) \quad (5)$$

The first camera pose x_0 can be given any value, since all other variables are estimated with relation to it, and is clamped to the origin in general.

In our measurements likelihood, we consider that the object position in image depends on the relative displacement between camera and object, and also on the camera orientation. The object shape is assumed independent of its position in image. Finally, we consider that the actual class labels C perfectly generates the detected ones $\bar{C}$, so that:

$$P(Z|X, M; J) = \prod_{k=1}^{K} \{P(u_k|x_{i_k}, l_{j_k})P(s_k|x_{i_k}, l_{j_k}, g_{j_k})\}$$

$$\delta\theta^* = \arg\min_{\delta\theta}\left\{\sum_{j=1}^{N}\left[||\delta l_j + l_j^0 - \gamma(c_j)||^2_{\Gamma(c_j)} + \frac{1}{2}||\delta g_j + g_j^0 - \varsigma(c_j)||^2_{\Sigma(c_j)}\right] + \sum_{i=1}^{T}||F_i^x\delta x_{i-1} - \delta x_i + f_i(x_{i-1}^0, v_i) - x_i^0||^2_{Q_i} + \sum_{k=1}^{K}\left[||U_k^x\delta x_{i_k} + U_k^l\delta l_{j_k} + h_k^u(x_{i_k}^0, l_{j_k}^0) - u_k||^2_{R_k} + ||S_k^x\delta x_{i_k} + S_k^l\delta l_{j_k} + S_k^g\delta g_{j_k} + h_k^s(x_{i_k}^0, l_{j_k}^0, g_{j_k}^0) - s_k||^2_{W_k}\right]\right\}$$

Table 1. Linear LS problem yielded by assuming linearized models for the odometers and the measurements. The superscript ^{0} indicates the linearization point of the respective variables, and δx means the variation of the variable x around its linearization point x^0, so that $x = x^0 + \delta x$. The decorated capital letters represent the Jacobians of the model functions: F_i^x is the Jacobian of $f_i(x, v)$ w.r.t x, U_k^x and U_k^l are, respectively, the Jacobians of $h_k^u(x, l)$ w.r.t. x and l, and S_k^x, S_k^l, and $S_k^g$are the Jacobians of $h_k^s(x, l, g)$ w.r.t. x, l, and g, respectively. The notation $||a||^2_\Sigma$ is used to indicate the squared Mahalanobis norm of a with respect to Σ, given by $a^T\Sigma^{-1}a$. Note that the variables subject to inference become $\delta\theta(\delta X, \delta L, \delta G)$, where $\delta X = \{\delta x_i\}_{i=1}^T$, $\delta L = \{\delta l_j\}_{j=1}^N$, and $\delta G = \{\delta g_j\}_{j=1}^N$.

As a result, the posterior in (1) is given by the generative model

$$\begin{aligned} P(X,M|Z;J) = P(x_0)\prod_{i=1}^{T}P(x_i|x_{i-1}, v_i) \\ \times\prod_{j=1}^{N}\{P(l_j|c_j)P(g_j|c_j)\} \\ \times\prod_{k=1}^{K}\{P(u_k|x_{i_k}, l_{j_k})P(s_k|x_{i_k}, l_{j_k}, g_{j_k})\} \end{aligned} \quad (6)$$

2.2. Assuming Simple Geometry: Size Only

In this work, we take g_j to be simply the object 3D bounding dimensions, and s_k the apparent size measurements. Although the generative model of the objects shape in images can be very complex, this simplifications can yield fairly approximated models under certain assumptions. The most important ones are that the camera keeps a certain distance from the objects, and that their apparent sizes do not change significantly from different viewpoints at the same distance.

The interesting difference with point-based visual SLAM or SFM is that apparent size now yields range to objects even by a single sighting. After several sightings both object dimensions and position will be sharply determined by triangulation, obsoleting the coarse priors.

3. Inference using QR decomposition

As inference technique, we adopt the same inference engine as many traditional SFM works. As usual in this literature, we factorize the posterior in (6) as product of Gaussian probabilities, which naturally leads (1) to be formulated as a linearized LS problem. In more complex 3D reconstruction (e. g., [11]), solving the linearized problem is part of an iteractive non-linear optimization strategy, like Levenberg-Marquardt. Here, we focus only the linear part.

3.1. Using linearized Gaussian models

To assure the posterior (6) is expressed as a product of Gaussians we define our model considering that all measurements and prior knowledge are normally distributed. Thus, the prior over objects location and size are given by

$$\begin{aligned} l_j &= \gamma(c_j) + e_j^l, \quad e_j^l \sim N(0, \Gamma(c_j)) \\ g_j &= \varsigma(c_j) + e_j^g, \quad e_j^g \sim N(0, \Sigma(c_j)) \end{aligned} \quad (7)$$

where e_j^l and e_j^g are the errors on the priors over objects location and size, respectively. Odometry and measurements are also disturbed by white noise, so we can write:

$$\begin{aligned} x_i &= f_i(x_{i-1}, v_i) + e_i^x, \quad e_i^x \sim N(0, Q_i) \\ u_k &= h_k^u(x_{i_k}, l_{j_k}) + e_k^u, \quad e_k^u \sim N(0, R_k) \\ s_k &= h_k^s(x_{i_k}, l_{j_k}, g_{j_k}) + e_k^s, \quad e_k^s \sim N(0, W_k) \end{aligned} \quad (8)$$

where e_i^x, e_k^u and e_k^s are, respectively, the odometry error, and the errors in the object position and size in image.

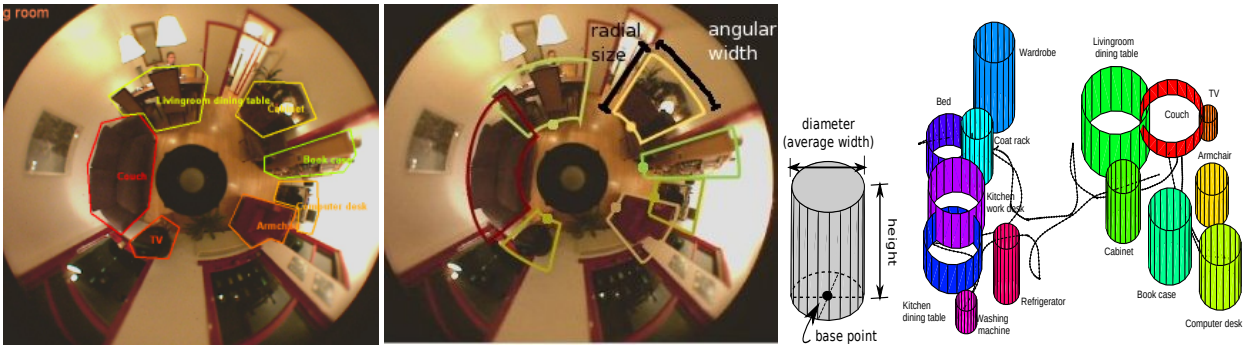
Since the functions $f_i(.)$, $h_k^u(.)$ and $h_k^s(.)$ are, in general, non-linear, linearized versions of them are used to assure a Gaussian posterior density. Replacing the densities yielded by (7) and by the linearized version of (8) in (6) yelds our posterior, so that taking the negative natural log of the maximizing term in (1) results in a linear LS problem, stated in table 1.

3.2. QR factorization

The resulting LS problem can be efficiently solved using a sparse Choleski factorization, like QR, by rewritting it in the matricial form:

$$\delta\theta^* = \arg\min_{\delta\Theta}||A\delta\theta - \mathbf{b}||^2_{\mathbb{P}} \quad (9)$$

where each block-line in A and $\mathbf{b}$ correspond to one of the summand terms in Table 1, and $\mathbb{P}$ is a block-diagonal matrix with the covariances that weigh the summands.



(a) (b) (c) (d)

Figure 4. Illustration of the data and object model used in our experiments. a) Example of annotated data available from the data set. Objects are marked with bounding polygons, and are given a label. c) Detected objects sizes in image, corresponding to a "bounding slice" of the object, extracted from the annotation polygon. They comprise the radial size, which is a projection of the object height, and the angular width. d) Model adopted to represent the objects, comprising their height, average width, and base point position. e) Exemple of the built 3D environment model, showing also the inferred trajectory.

Due to the sparseness of A, QR factorization is an efficient way to solve (9) [12]. QR factorization represents an $m \times n$ matrix A, with $m \geq n$, by a multiplication of other two matrices [8], $A = Q \begin{bmatrix} R \\ 0 \end{bmatrix}$, where Q is an $m \times m$ orthonormal matrix, and R is the $n \times n$ upper-triangular Cholesky factor of $A^T A$. Let us rewrite (9) as a minimization of an Euclidean distance by incorporating the covariance matrix into the other terms:

$$||A\delta\theta - \mathbf{b}||^2_{\mathbb{P}} = ||\bar{A}\delta\theta - \bar{\mathbf{b}}||^2$$

where $\bar{A} = \mathbb{P}^{-\frac{1}{2}} A$, $\bar{\mathbf{b}} = \mathbb{P}^{-\frac{1}{2}} \mathbf{b}$, and $\mathbb{P}^{-\frac{1}{2}}$ is the Cholesky factor of $\mathbb{P}^{-1}$. The QR factorization of $\bar{A}$ allows us to rewrite (9) in the form

$$\begin{aligned} \delta\theta^* &= \arg\min_{\delta\Theta} ||Q \begin{bmatrix} R \\ 0 \end{bmatrix} \delta\theta - \bar{\mathbf{b}}||^2 \\ &= \arg\min_{\delta\Theta} ||Q \left(\begin{bmatrix} R \\ 0 \end{bmatrix} \delta\theta - \begin{bmatrix} \mathbf{c} \\ \mathbf{r} \end{bmatrix} \right)||^2 \end{aligned}$$

Once $\bar{R}$ is upper-triangular square, and full-rank since $\bar{A}$ poses an over-determined linear system, the solution for the problem is obvious: it is given by the solution of $R\delta\theta = \mathbf{c}$, leaving $||\mathbf{r}||^2$ as the total squared residual.

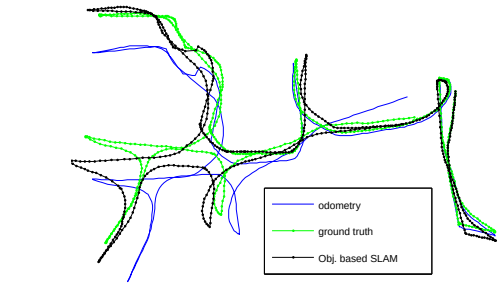
4. Experimental results

The presented approach was tested using a real-world annotated data set, obtained on line from the project *From Sensors to Human Spatial Concept* [13] website. The image sequence was captured by an omnidirectional camera using a hyperbolic mirror at 7.5 fps. However, only the odd numbered images were annotated, so we use images grabbed at half of this frame rate. We assume that the objects in the scene could be roughly represented by cylinders, using the 3D position of their base point to represent their locations, as depicted in Fig. 4. The measured sizes correspond to the angular slice bounding the annotated object. We clearly benefit from using an omnidirectional camera, once we can get several sights of an object without concerning about actively focusing it in the image. For the camera movement, we assumed a 3DOF planar motion model, using odometer information to predict the camera pose from a frame to the other.

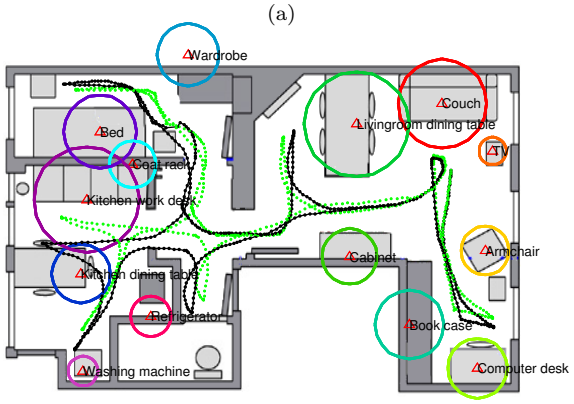
The linearization value of the object parameters were initialized by projecting their *a priori* height in the world using the radial size measurement. Obtained results are showed in Figs. 4(d) and 5. Our approach showed an improvement of the recovered camera path w.r.t. the *a priori* odometer-based one. As a ground truth for our experiments, we used the trajectory obtained by a SLAM algorithm using laser scans data provided in toolbox that came with the data set.

5. Conclusion and future work

We presented a novel approach for acquiring simple 3D object-based models of the environment from a single moving camera. The results presented corroborate the idea of using object recognition output in a simple and fast 3D model builder of the objects in the scene. More elaborated models would require view-point dependent modeling, leading to inference on a hybrid discrete/continuous model, which are more costly to infer.



(a)



(b)

Figure 5. Results obtained with object-based SLAM. a) Comparison among trajectories. The inferred trajectory is closer to the ground truth. b) Bird-eye view of the built objects model. The black-dotted line represents the inferred trajectory, while the green line is the trajectory obtained by laser-based SLAM, which we consider as our ground truth. The house blueprint was extracted from [13].

Although we require objects to be recognized perfectly for now, state-of-the art object recognition techniques show very low error rates, so that a quasi-ideal condition could be achieved indoors, by focusing on very prominent obstacles, and tuning the detector to minimize fake positives. However, for future work we are working on augmenting our model to include discrete unknowns, namely the correspondence J, the class labels C, and possibly viewpoint selection. More specifically, we are investigating sampling techniques over the discrete parameter space, each sample corresponding to a different linear system. In this case, we will investigate incremental QR update techniques, so that small changes in the discrete variables can be rapidly processed, leading to vary fast estimates.

References

[1] S. Vasudevan, S. Gachter, M. Berger, and R. Siegwart, "Cognitive maps for mobile robots — an object based approach," *Journal of Robotics and Autonomous Systems*, vol. 55, pp. 359–371, May 2007.

[2] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J. Fernández-Madrigal, and J. G. ez, "Multi-hierarchical semantic maps for mobile robotics," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 3492–3497, 2005.

[3] B. Williams, G. Klein, and I. Reid, "Real-time SLAM relocalisation," in *Intl. Conf. on Computer Vision (ICCV)*, 2007.

[4] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *IEEE and ACM Intl. Sym. on Mixed and Augmented Reality (ISMAR)*, (Nara, Japan), November 2007.

[5] A. Davison, I. Reid, N. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 6, pp. 1052–1067, 2007.

[6] R. O. Castle, D. J. Gawley, G. Klein, and D. W. Murray, "Towards simultaneous recognition, localization and mapping for hand-held and wearable cameras," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, (Rome, Italy), pp. 4102–4107, Apr. 2007.

[7] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2137–2144, 2006.

[8] G. Golub and C. V. Loan, *Matrix Computations*. Baltimore: Johns Hopkins University Press, third ed., 1996.

[9] F. Dellaert, "Square Root SAM: Simultaneous location and mapping via square root information smoothing," in *Robotics: Science and Systems (RSS)*, 2005.

[10] M. Kaess, A. Ranganathan, and F. Dellaert, "Fast incremental square root information smoothing," in *Intl. Joint Conf. on AI (IJCAI)*, (Hyderabad, India), pp. 2129–2134, 2007.

[11] K. Ni, D. Steedly, and F. Dellaert, "Out-of-core bundle adjustment for large-scale 3D reconstruction," in *Intl. Conf. on Computer Vision (ICCV)*, (Rio de Janeiro), October 2007.

[12] F. Dellaert, "Square root SAM," in *Proc. of Robotics: Science and Systems*, (Cambridge, MA), June 2005.

[13] Z. Zivkovic, O. Booij, and B. Kröse, "From images to rooms," *Journal of Robotics and Autonomous Systems*, vol. 55, no. 5, pp. 411–418, 2007. Data set url: `http://staff.science.uva.nl/~zivkovic/FS2HSC/dataset.html`.