Automated Multisensor Polyhedral Model Acquisition

D. Ortín [1] J.M.M. Montiel [1]
[1] Computer Science and Systems Engineering Dept.
University of Zaragoza.
María de Luna 1, 50018 Zaragoza. Spain.
{dortin,josemari}@unizar.es

A. Zisserman [2]
[2] Engineering Science Dept.
Oxford University.
Parks Road 19. OX1 3PJ. Oxford. U.K.
az@robots.ox.ac.uk

***Abstract*— We describe a method for automatically generating accurate piecewise planar models of indoor scenes using a combination of a 2D laser scanner and a camera on a mobile platform.**

The method exploits the complementarity of the sensors. Mapping techniques applied to 2D laser scans simultaneously compute a map and the location of the sensor in the unknown environment. This provides an initial estimate for the vision algorithms by compensating for rotation, foreshortening and scale change between images. The vision algorithms are then able to compute a very accurate registration (via a plane to plane homography) which is used to segment the model into planar facets, and to improve the estimate of the model and sensor position.

Results are demonstrated on a man made scene using a 2D laser scanner and a calibrated camera mounted on a trolley.

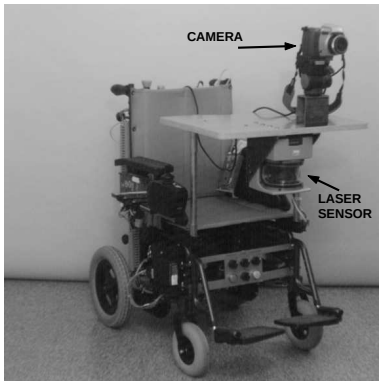

Fig. 1. Sensorized trolley. In the experiments we use the SICK 2D laser scanner and the vision camera. The camera is a 1280×1024 photogrammetric calibrated camera.

I. Introduction

We focus on two approaches to the problem of simultaneous estimation of a sensor motion and the observed scene structure from sensor input. The first is multisensor SLAM (Simultaneous sensor Location And Map building). The second is uncalibrated computer vision reconstruction. Our goal is the cross-fertilization between these two approaches in order to model and detect facades of indoor scenes.

The robustness of computer vision algorithms is improved if an initial estimation of the solution is available. We propose to use SLAM, with non-vision sensors, to provide it. On the other hand, vision can provide quite accurate angular measurements and additional redundancy, and SLAM can benefit from them. This cooperation in terms of robustness and accuracy is demonstrated in this paper.

In detail we aim to detect planar elements (facades) in indoor man made environments using sensors mounted on a trolley (actually a wheelchair). The trolley (see fig. 1) has a calibrated semimetric camera and a 2D laser, but no odometry is used. A hand-held tour is done around a room, gathering continuously laser scans (at 1 per second rate), and taking manually several pictures at some key locations. The goal is to automatically process all the gathered data.

First, a 2D map of the area and the location of the trolley in it is computed using SLAM, and then the map is used as an initial guess for computer vision processing, upgrading 2D lines to 3D planes. Automatic matches for points on the planes and a homography mapping are computed by the vision algorithms and used to delineate the planes. Finally, a photogrammetric reconstruction of both scene and robot locations is produced. An alignment for the images of one pixel accuracy is achieved. The outcome is a 3D reconstruction, with photogrammetry used to improve on the initial SLAM estimates.

A. Background

A review of uncalibrated computer vision can be found in [5] and [8]. We are interested in robust matching, a combination of projective geometry, image processing and robust statistics that has produced algorithms able to cope with real images under non-lab conditions; a review of these techniques can be found also in [8].

SLAM, developed mainly in the robotics community, exploits the complementarity of the information provided by different sensors mainly: dead reckoning, laser range finder, sonar and calibrated vision detecting discrete features. The information provided by all the sensors is reduced to a common Euclidean geometric framework in order to be combined. The feature locations are represented using a stochastic map. The stochastic map, proposed initially by Smith and Cheeseman in [13] is the central concept in SLAM. Since then several implementations have been successfully reported, especially relevant for us are [3], [2] and [6] for

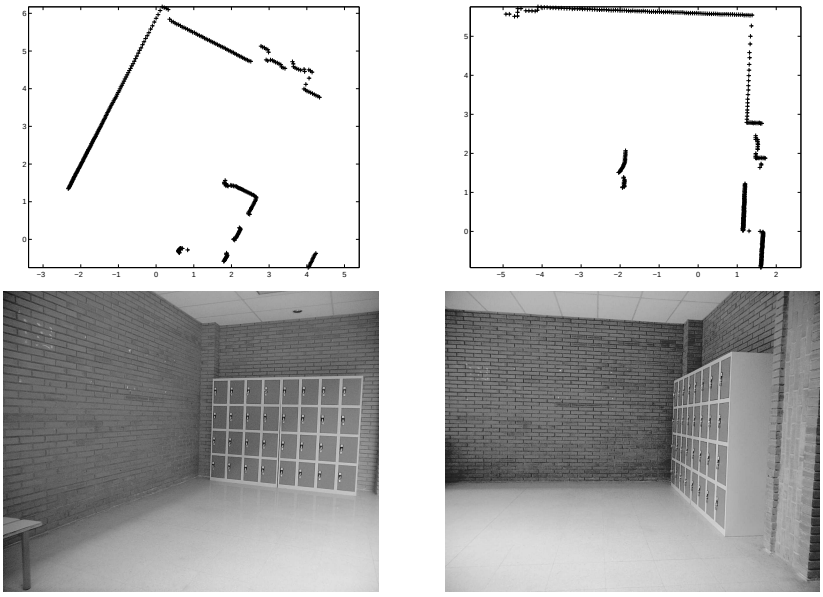

Fig. 2. Top shows 2 sample laser scans. The corresponding images below were acquired from the same trolley location, illustrating the structure of the environment. These images are also used for plane detection. Notice the different scale and foreshortening for the brick wall and for the lockers.

multisensor fusion using discrete features, and [4] for real time active vision. The correlation among all the features in the map has proven to be an essential factor to deal with error drift, produce accurate error estimates, and manage realistic size maps. In this work the use of joint compatibility data association [10] allows a laser SLAM to be computed without any odometric information.

Geometrically, the computer vision task dealt with in this work is similar to mosaicing images from a rotating camera or a planar surface [1], [9]. Both cases are modelled by a homography. In the rotation case, the (image of) the entire rigid scene may be mapped by the same homography, while in the case that the camera translates (as well as rotates) only the image of points on a scene plane are mapped by a homography, and thus segmentation of the image into planar regions is necessary. In our work, pre-segmentation is aided by non visual information about the scene and camera motion.

The present work is also related to wide baseline matching [11] where feature matches are used to determine initial homographies between image pairs. Image synthesis (by homography warping or more elaborate methods) has also been used by [7] to improve the correlation based matching performance in the field of video processing for accurate video insertion and video annotation.

In the literature a number of devices are described for automated man-made model acquisition [12],[14],[15],[16]. Sequeira *et. al* [12] use a 3D range finder to detect the scene structure and a color video camera to add the texture. In our work we use a 2D laser scanner, but the cameras are used not only to detect the texture but also to infer the scene structure. Teller, in [14] proposes a system based on omnidirectional images to recover urban scenes. The omnidirectional images are synthesized from a rotating camera, and their wide angle improves the conditioning of the computer vision problem. The system can deal with huge scene sizes, using mainly visual information. Taylor [15] uses also an omnidirectional camera to locate a robot with respect to known landmarks.

II. Laser 2D map

We will only give a sketch here of the SLAM laser processing, a detailed description can be found in [3]. The trolley gathers a laser scan per second continuously. Fig. 2 shows two sample scans. The laser points are segmented into a set of straight segments.

The laser scans are processed sequentially in a prediction-match-update loop. In every step there is a data association stage to match the new laser segments with the available map, non matched segments are included as new map features; then the map is reestimated considering the new matches. Classical data association matches every segment as an isolated entity, however in this work we have used a joint compatibility data association [10] that considers jointly all the matches; this makes the data association robust and the map can be built without using any odometric information. Fig. 3 shows the final 2D map.

III. Two view plane geometry

The geometry of a plane observed by two cameras (see fig. 4), plays a central role in this work, and we now review the inter-image relations induced by a scene plane.

Camera location is defined by a frame attached to its optical centre. Let $\boldsymbol{X} = (X, Y, Z, 1)^\top$ be the homogeneous

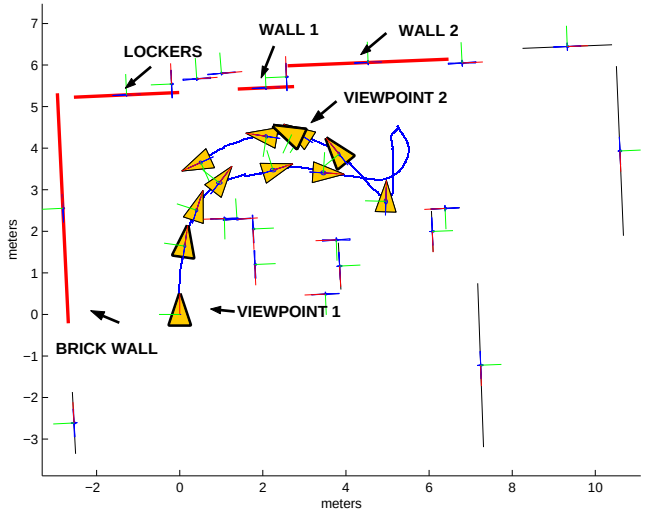

Fig. 3. The 2D map computed from the laser range finder data without the use of any odometry. The trolley trajectory is shown as well. The locations where an image was acquired are shown as triangles. The two trolley locations (see the corresponding images in fig 2) and the two planes (brick wall and lockers) used in the example are marked.

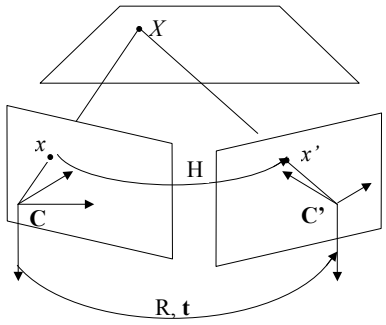

Fig. 4. Two view geometry of a plane.

coordinates of a 3D point with respect to the camera $\boldsymbol{C}$. Let elements with $\prime$ denote elements in the second camera. The second camera location is represented by the change of coordinates $\boldsymbol{R}$ and $\boldsymbol{t}$ from $\boldsymbol{C}'$ to $\boldsymbol{C}$.

If the location of an image point (u, v) is represented by its homogeneous coordinates: $\boldsymbol{x} = \lambda(u, v, 1)^\top \quad \forall \lambda \neq 0$, then the two images of the same point $\boldsymbol{X}$ on the plane are related by 3×3 homography $\boldsymbol{H}$ (see [8]):

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{H}\boldsymbol{x}' \\ \boldsymbol{H} &= \lambda \boldsymbol{K} \left(\boldsymbol{R} - \frac{\boldsymbol{t}\boldsymbol{n}^\top}{d} \right) \boldsymbol{K}'^{-1} \quad \forall \lambda \neq 0 \end{aligned} \tag{1}$$

where $\boldsymbol{K}$ is the camera calibration matrix [8]. The plane equation in the $\boldsymbol{C}$ frame is $(\boldsymbol{n}\ d)^\top \boldsymbol{X} = 0$.

IV. Scene plane instantiation and mapping

This section describes the algorithm for facade mapping starting from the 2D laser SLAM map and two images of the facade.

An initial guess for both camera location and polyhedral structure is derived from the laser map. A 2D laser scan segment instantiates a vertical wall plane (a facade) of the same width as the segment, and with a nominal height for the room.

Given two views, each rectangular facade projects onto the images as two areas of interest. Corresponding points in those areas are related by the mapping of (1). The images are prealigned according to this homography using the SLAM estimates of rotation and translation.

Improved alignments are then used to correct the relative camera-planes location. Although these homographies can be found using point correspondences the automatic computation between different views is prone to error as different features may be computed in each image. Therefore, a robust fitting method able to deal with spurious matches must be used. A detailed description of the classical robust estimation of an homography can be found in [8]. The algorithm is sketched in fig. 5. We include the additional step (1) to exploit the initial guess for the homography derived from the SLAM map.

The algorithm has a random nature and has a probability of succeeding p, related to the number of attempts at step (3) according to: $N = \frac{\log(1-p)}{\log\left(1-(1-\epsilon)^4\right)}$ where $\epsilon = \frac{\text{number of spurious matches}}{\text{number of putative matches}}$ is the spurious (outlier) rate.

N can be adaptively computed. A high spurious rate is initially assumed. When step (4,ii) computes the number of inliers, it sets an upper bound on the number of outliers and therefore on N.

The previous algorithm works well provided steps (2) and (3) produce a set of mostly correct putative matches. This can be done using correlation if the plane is detected in both images with a similar scale, rotation and foreshortening. We are dealing with images that do not fulfill these conditions: for example, the two images of the lockers in fig. 2. However, the prealignment of the images using the motion and scene initial guess corrects for these significant perspective deformations.

The algorithm is illustrated for the images of Fig. 2. Fig. 6 shows the various steps in the processing.

Input data :
2 images of the plane
Estimated location for camera and plane
Output data:
Final Homography $\boldsymbol{H}$
Point Matches $\{\boldsymbol{x}_i, \boldsymbol{x}'_i\} \quad i = 1 \dots n$
Algorithm
(1) Initial image alignment : from SLAM
(2) Interest point detection: Harris detector
(3) Putative correspondences: compute point matches based on image proximity and similarity of their intensity neighbourhood. Similarity is measured with correlation
(4) RANSAC $\boldsymbol{H}$ estimation:
Repeat for N attempts:
(i) Select randomly 4 point matches. Compute $\boldsymbol{H}$
(ii) Calculate how many putative matches are inliers, i.e are consistent with $\boldsymbol{H}$
Select the $\boldsymbol{H}$ with most inliers.
(5) Nonlinear $\boldsymbol{H}$ estimation: from the inliers
(6) Guided matching: Further matches are determined using $\boldsymbol{H}$ to define the search region.
Steps (5) and (6) are iterated until the correspondences are stable

Fig. 5. Planar homography computation algorithm.

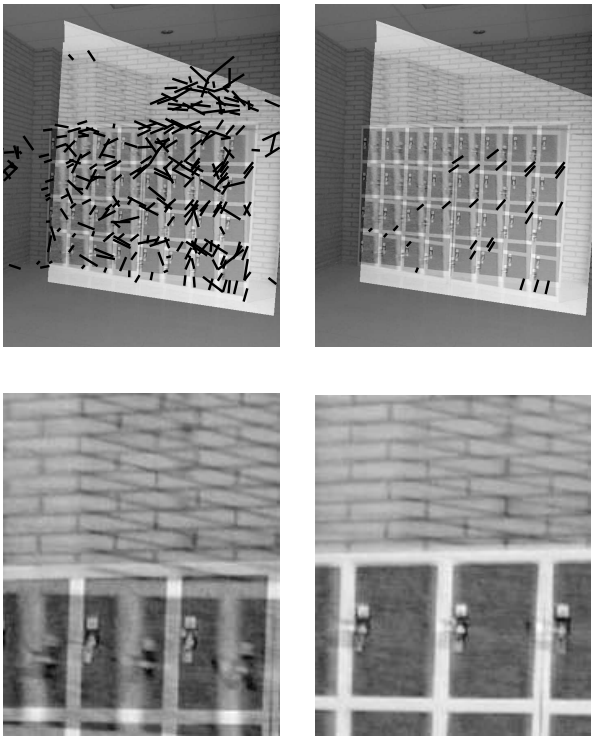

Fig. 6. Top row shows the images in fig. 2 aligned according to the SLAM map for the lockers plane. The considered region of interest is highlighted. Top left image shows the putative matches. Top right the inliers, all of them in the plane of the lockers. The bottom row shows a detail of the initial SLAM image alignment (left) and the alignment with the computed homography (right); the final alignment for the plane of the lockers is accurate up to the pixel level.

V. Plane Segmentation

Having computed the homography for the plane we now have a point to point map available between the images. The goal then is to determine which image pixels are image of a facade. The idea is to segment the plane by identifying which pixels are consistent with the supplied homography mapping. The input data consists of two views of a scene plane and the plane induced homography.

If the surfaces were Lambertian and the illumination constant, two images of the same facade aligned according to the computed homography should be coincident for every pixel. To increase the robustness with respect to illumination changes, a window around every pixel is used to determine the similarity between the pixels, and thereby if the pixel is an image of the considered plane.

The score used is the normalized correlation over the window around each pixel after having aligned the images according the computed homography.

$$NCC = \frac{\sum_{i,j} \left(\boldsymbol{w}_{ij} - \bar{\boldsymbol{w}}\right)\left(\boldsymbol{w}'_{ij} - \bar{\boldsymbol{w}'}\right)}{\sqrt{\sum_{i,j} \left(\boldsymbol{w}_{ij} - \bar{\boldsymbol{w}}\right)^2 \sum_{i,j} \left(\boldsymbol{w}'_{ij} - \bar{\boldsymbol{w}}'\right)^2}}$$

where $\boldsymbol{w}_{ij}$ and $\boldsymbol{w}'_{ij}$ are the intensities for two windows centred respectively around the corresponding aligned pixels. Isolated plane regions smaller than 0.5% of the image surface are removed. Fig. 7 shows the segmented plane for the lockers' facade.

VI. Bundle Adjustment from point matches

Once the point matches have been computed, a conventional bundle adjustment can be applied to recover both the 3D scene and the cameras' location. It consists of a non linear optimization to reduce the re-projection error in the image.

Let $\boldsymbol{X}_j$ be the scene points, and P^i the calibrated cameras (whose only free parameters are their location). Then $P^i \boldsymbol{X}_j$ is the theoretical image of point $\boldsymbol{X}_j$ seen by camera P^i actually detected as point $\boldsymbol{x}^i_j$.

We want to compute the camera motion P^i and 3D point locations $\boldsymbol{X}_j$ such that:

$$\sum_{i,j} d\left(P^i \boldsymbol{X}_j, \boldsymbol{x}^i_j\right)^2$$

is minimized, where $d\left(P^i \boldsymbol{X}_j, \boldsymbol{x}^i_j\right)$ is the geometric distance between the predicted and the actual image of a 3D point. See [8] for further details.

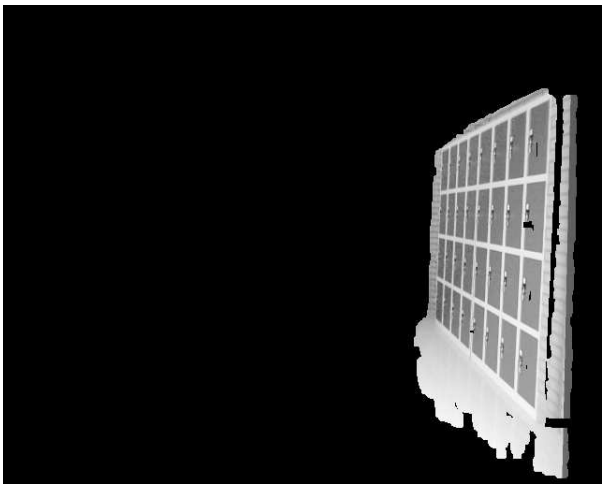

Fig. 7. Segmented image area as belonging to the plane of the lockers.

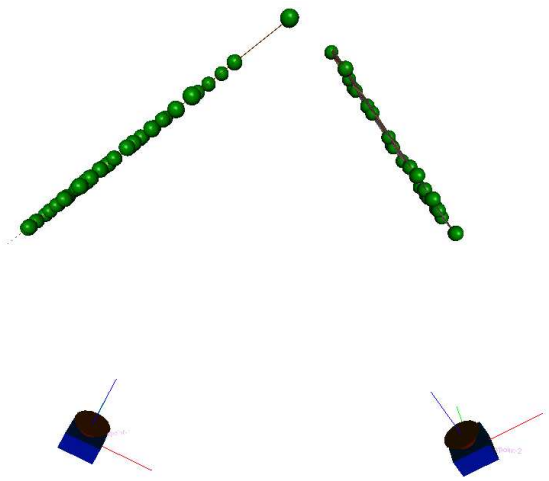

Fig. 8. 3D reconstruction after bundle adjustment. The 3D location for the matched points in the two views. The detected points are a very good fit to the two planes corresponding to the corner.

TABLE I

RANSAC RESULTS FOR ALL THE CORRECT PVPS

	mean	min	max
inliers fraction	39%	18%	68%
RANSAC iterations N	524	34	4479
SLAM alignment error (px)	31.7	2.8	70.7
final alignment error (px)	1.0	0.5	1.7

Fig. 8 shows the points and the cameras after the bundle adjustment, for the brick wall and lockers facade in two views. Fig. 9 shows the results of these two planes in 4 views.

VII. EXPERIMENTAL RESULTS

To test the performance of the proposed algorithm in detecting planes, it has been applied to the four planes marked in the map shown in fig. 3. The goal was to compute automatically all the pairwise matches and from them to compute a bundle adjustment.

We call PVP (Plane View Pair) a plane detected in two different views. Considering the four planes over all the views, there were 39 feasible PVPs. The goal was to automatically detect two view point matches belonging to the plane. After manual verification the results were:

- 3 PVPs were not detected because less than 8 putative matches were found. They corresponded to planes with small overlap between the views.
- 1 PVP was detected erroneously. However only 10 inlier matches were detected, so if only PVPs with a high number of matches were accepted the system could automatically detect only correct PVPs.
- 35 PVPs were correctly detected.

Table I summarizes the results of the robust RANSAC plane estimates. Due to the high spurious rate it is absolutely necessary to use robust statistics. It can be seen how the number of inliers is sometimes under 50%; the algorithm copes with this because the outliers error is uncorrelated and there is not a model consistent with all of them.

The outliers fraction cannot be anticipated so it is efficient to estimate it adaptively, as detailed in section IV.

The computed homography was able to reduce the misalignment of the SLAM initial guess, whose initial alignment errors were unacceptable for computer vision applications.

VIII. DISCUSSION AND FUTURE WORK

The automatic acquisition of indoor man-made polyhedral models using a sensorized trolley has been shown to be possible. The key idea has been the combination of a laser 2D range finder and a calibrated camera. An accuracy in image alignment of about one pixel has been achieved.

The results show the feasibility of this technique with real data, however more work is needed for reliable detection and mosaicing of plane textures. For example, although the plane delineation works well in textured areas, it is ambiguous in untextured areas.

Future work is aimed at integrating the visual information into the SLAM formalism for the joint consideration of the vision and laser constraints.

ACKNOWLEDGEMENTS

This work has been partially supported by the Spanish Government PR2001-0149, MCYT DPI 2000-1265 and the CAI CONSI+D (Prog. Europa Ref IT/401).

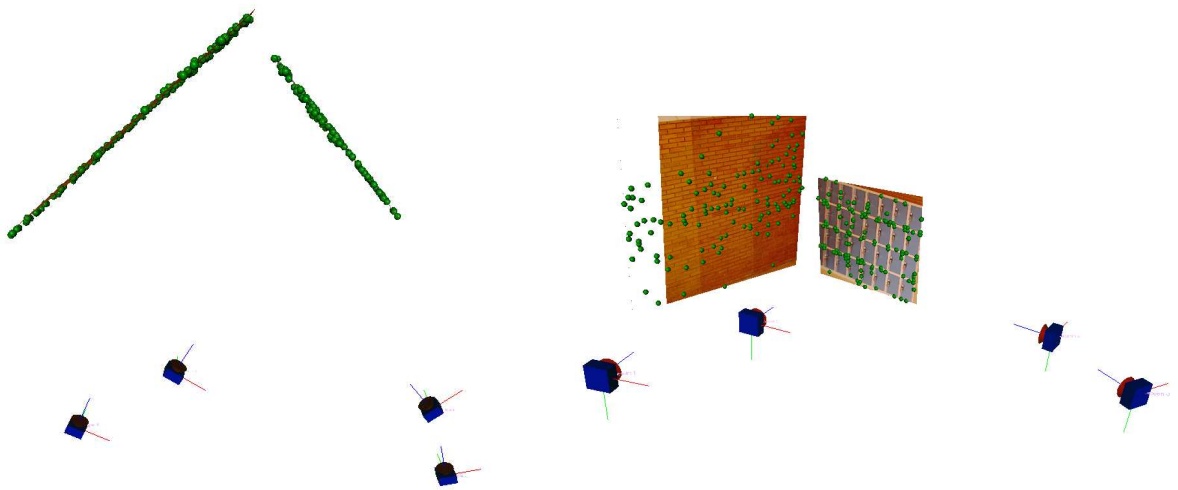

Fig. 9. 3D reconstruction from point matches. Left shows a top view and right a general one.

REFERENCES

[1] D. Capel and A. Zisserman. Automated mosaicing with super-resolution zoom. In *Proc. CVPR*, pages 885–891, Jun 1998.

[2] J.A. Castellanos, J.M.M. Montiel, J. Neira, and J.D. Tardós. Sensor influence in the performance of simultaneous mobile robot localization and map building. In P. Corke and J. Trevelyan, editors, *Experimental Robotics VI. Lecture Notes in Control and Information Sciences. Vol 250*, pages 287 – 296. Springer-Verlag, 1994.

[3] J.A. Castellanos and J.D. Tardós. *Mobile Robot Localization and Map Building: A Multisensor Fusion Approach*. Kluwer Academic Publishers, Boston. USA, 1999.

[4] A.J. Davision and D.W. Murray. Simultaneous localization and map building using active vision. *IEEE Trans. on PAMI*, 24(7):865 – 880, July 2002.

[5] O.D. Faugeras, Q. T. Luong, and T. Papadopoulo. *The Geometry of Multiple Images*. MIT Press, 2001.

[6] H.J.S. Feder, J.J. Leonard, and C.M. Smith. Adaptive mobile robot navigation and mapping. *Int. Journal of Robotics Research*, 18(7):650–668, 1999.

[7] K.J. Hanna, H.S. Sawhney, R. Kumar, Y. Guo, and S. Samarasekara Annotation of video by alignment to reference imagery. In Zisserman A. Triggs, B. and Szeliski R., editors, *Vision Algorithms'99 LNCS 1883*. Springer-Verlag, 2000.

[8] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.

[9] M. Irani, P. Anandan, and S. Hsu. Mosaic based representations of video sequences and their applications. In *5th ICCV, Boston*, pages 605–611, 1995.

[10] J. Neira and J. D. Tardos. Data association in stochastic mapping using the joint compatibility test. *IEEE Trans. R&A*, 17(6):890 – 897, Dec 2001.

[11] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *6th ICCV, Bombay*, pages 754–760, Jan 1998.

[12] V. Sequeira, K. Ng, E. Wolfart, J.G.M. Gonzalvez, and D.C. Hogg. Automated reconstruction of 3d models from real environments. *ISPRS Journal of Photogrammetry and Remote Sensing*, 54:1–22, 1999.

[13] R. C. Smith and P. Cheeseman. On the representation and estimation of spatial uncertainty. *Int. J. Robotics Research*, 5(4):56–68, 1986.

[14] S. Teller. Scalable, controlled image capture in urban environments. Technical report, Technical Report 825. MIT Lab CS., September 2001.

[15] C.J. Taylor. Videoplus: A method for capturing the structure and appearance of immersive environments. *IEEE Transactions on Visualization and Computer Graphics*, 8(2):171–182, Apr-Jun 2002.

[16] Yufeng Liu, R. Emery, D. Chakrabarti, W. Bugard, and S. Thrun. Using EM to learn 3d models of indoor environments with mobile robots. In *International Conference on Machine Learning*, pages 329–336, Williams College, Williamstown, MA, USA, June-July 2001.