Landmark-Based Navigation – Model Extension and Refinement

Harpreet S. Sawhney
Rakesh Kumar
Allen R. Hanson
Edward M. Riseman

COINS TR93-06

January 1993

LANDMARK–BASED NAVIGATION — MODEL EXTENSION AND REFINEMENT [1]

Harpreet S. Sawhney Rakesh Kumar
Allen R. Hanson Edward M. Riseman

Computer Science Department
University of Massachusetts
Amherst, MA 01003
Phone : (413)545–2744

Abstract

In [5], least-squares and robust methods were presented for determining the location and orientation of a mobile robot from visual measurements of modeled 3D landmarks. However, building the 3D landmark models is a time consuming and tedious process. For landmark-based navigation methods to be widely applicable, automatic methods have to be developed to build new 3D models and enhance the existing models. Ideally, a robot would continuously build and update its world model as it explores the environment. This paper presents techniques to determine the 3D location of image features from a sequence of monocular 2D images captured by a camera mounted on the robot.

The approach adopted here is to first build a partial model (possibly noisy) either manually, by stereo, or by tracking and reconstructing *shallow* structures over a sequence of images using the constraint of affine trackability. This model is subsequently used to compute the pose that relates the model coordinate system and the camera coordinate system of the image frames in the sequence. The unmodeled 3D features (those not already in the model) are tracked over the image sequence and their 3D locations recovered by a pseudo–triangulation process, a form of "induced stereo". The triangulation process is also used to make new 3D measurements of the initial model points. These measurements are then fused with the previous estimates to refine the set of initial model points.

1 Introduction

In [5], least-squares and robust methods were presented for determining the location and orientation of a mobile robot from visual measurements of modeled 3D landmarks. However, building the 3D landmark models is a time consuming and tedious process. For landmark-based navigation methods to be widely applicable, automatic methods have to be developed to build new 3D models and enhance the existing models. Ideally, a robot would continuously build and update its world model as it explores the environment. This paper presents techniques to determine the 3D location of image features from a sequence of monocular 2D images captured by a camera mounted on the robot. The approach adopted here is to first build a partial model (possibly noisy) and to then extend and refine it by viewing the scene over a sequence of frames. The partial model is derived from the reconstruction of shallow[2] environmental structure [10]. Model extension results are also presented for one sequence where the partial model was manually built.

1.1 Related Work

Previous research on multi-frame 3D reconstruction can be categorised into two broad classes. The first class assumes that a model of 3D inter-frame motion is known, rather than assuming independent motion parameters between consecutive frames. Broida [2] assumes constant velocity motion and estimates the 3D location of a set of points tracked over a monocular image sequence. Chandrasekhar et. al. [3] have extended Broida's technique to deal with data sets where the 3D location of a few points is known. The objective function, which Broida and Chandrasekhar et. al. minimize, has the motion model parameters and the unknown structure location parameters

[1] This work was supported in part by DARPA (via TACOM) under contract number DAAE07–91–C–R035, and by NSF under grant number CDA–9822572.

[2] Shallow structures have small extent in depth compared to their distance from the camera.

as unknowns. Thus the dimension of the objective function grows with the number of unknown points. An even more basic limitation of this approach lies in the model of motion being adopted and its suitability to the motion being observed.

The second class of techniques does not assume any model of motion. The rigid structure of the world is carried forward by the depth estimates from frame to frame. These techniques are sequential in nature and typically use Kalman Filtering to compute the depth estimates[4, 8, 10]. Oliensis and Thomas [8] solve for the motion parameters between consecutive image frames in a monocular image sequence. With each image pair, new measurements are made for depth values of features and these are integrated with previous estimates in the Kalman Filter framework. The new observation Oliensis and Thomas [8] make is that the depth estimate of different feature points are correlated since the same noisy motion parameters are used to compute the depth. Because of this correlation, they estimate the depth parameters of all points simultaneously. This gives them fairly good depth estimates for camera motions having some T_z (i.e. translation along the optical axis) component. The cost, however, is that for estimating the depths of m points, a covariance matrix of size $3m \times 3m$ must be inverted with each new frame.

1.2 Overview

All of these approaches rely on the basic principle of triangulation to reconstruct new 3D points. However, reconstruction by triangulation is highly sensitive to errors in estimating the relative orientation between consecutive camera frames. In this paper, the reconstruction of 3D structure is accomplished in two steps to overcome this limitation.

The first step is a partial reconstruction of a scene in terms of *shallow* 3D environmental structure; structures whose extent in depth is small compared to their distance from the camera. The 3D motion and structure of a shallow object in motion, relative to the camera, can be well approximated by an affine transformation. In [10], a framework was presented for tracking shallow objects over time under the affine constraint, and in [11] an algorithm for identification and 3D reconstruction of these structures is presented. An important advantage of this approach is that 3D structure is derived reliably without the intermediate step of explicit computation of the 3D motion parameters.

Shallow structure reconstruction provides only a partial 3D model for the scene. However, this partial model is adequate for the second part of the technique presented in this work, namely model extension and refinement. The partial model is used to compute the pose that relates the model coordinate system and the camera coordinate system of the image frames in the sequence. The unmodeled 3D features (those not recovered by the shallow structure reconstruction) are tracked over the image sequence using an optic flow based line tracking algorithm [13]. Using correspondences of image features, and the poses computed from model–to–image feature correspondences for a sequence, new 3D points are located by triangulation (see Figure 1). The estimation of the new 3D points is done using both batch and quasi-batch or sequential methods. The triangulation process is also used to make new 3D measurements of the initial model points which are then fused with the previous estimates to refine the set of initial model points. Results are presented for real sequences where new 3D points are located with average errors less than 1.76 %.

Note that this approach does not require any models of inter-frame motion. Due to the availability of the partial model, new points are located in a stable world coordinate system. The pose computed for each frame is independent of the other frames, so each frame provides an independent measurement to the whole process[3]. This does not lead to the cascading problems which most of the sequential multi-frame "structure from motion" techniques suffer from because, in the latter, noisy prior estimates of motion in the previous frame are used to propagate the structure estimates to the next frame which are then integrated with the new estimates in the current frame.

The errors in the initial partial model (for the model extension and refinement step) are assumed to be either gross errors or gaussian noise. If gross errors are present in the 3D model, these would be detected as outliers by the robust pose recovery techniques developed earlier [5] and would not be used for the final step of least-squares fitting to the remaining non-outlier data. Note that outliers can also arise due to incorrect correspondences. However, if

[3]Note that this would not be true if there was significant noise in the initial partial model.

a modeled landmark appears as an outlier over a large number of frames, then it probably is due to a gross error in the 3D model and it could eventually be removed from the 3D model database. Thus, for the remainder of this paper, the noise in the input 3D model is modeled as gaussian.

2 Shallow Structure Reconstruction

This section presents a brief summary of identification, tracking and 3D reconstruction of shallow structures. The details can be found in [11] in this proceedings.

Given a 3D structure that can be well approximated by a fronto–parallel plane (shallow structure), its image projections at two closely spaced time instants are related through:

$$\frac{1}{f}p' \approx \frac{1}{f}sR_z p + t, \quad t = s\Omega_{xy} + \frac{1}{Z'_0}T_{xy} \tag{1}$$

where, p and p' are the corresponding imaged points of a shallow structure at times n and $n+1$ respectively, s is the scale defined as the ratio of average depths at the two time instants, R_z is the 2×2 rotation matrix for the rotation around the optical axis (z–axis), t is the translation in the image plane, Ω_{xy} and T_{xy} are the vectors representing the x and y components of the 3D rotational and translational vectors respectively, Z'_0 is the average depth at the second time instant, and f is the focal length of the camera.

A set of noisy line correspondences are used to compute the best affine motion parameters in the image plane. An error measure that is a weighted sum of the parallel and perpendicular components of the vectors joining the corresponding endpoints of the line in frame $n+1$ and the affine transformed line in frame n is formulated:

$$E_i = \sum_{j=1}^{2} w_{\perp_i}[(D_{ij}r_s + t - p'_{ij}) \cdot n'_i]^2 + w_{\parallel_i}[(D_{ij}r_s + t - p'_{ij}) \cdot l'_i]^2 \tag{2}$$

where i is the ith corresponding pair, j refers to endpoint 1 or 2, $w_{\perp_i}$ and $w_{\parallel_i}$ are the weights for the perpendicular and parallel error components, $D = \begin{bmatrix} x & -y \\ y & x \end{bmatrix}$ is the data matrix which is constructed using the endpoint $p = [x \;\; y]^T$ in frame n, vector $r_s = [s\cos\omega_z \;\; s\sin\omega_z]^T$ is the product of scale s and rotation, ω_z, around the optical axis, and n'_i and l'_i are the unit normal and direction, respectively, of the line in frame $n+1$.

For a set of line correspondences, the unknown parameters r_s and t can be found by minimising $\sum_i E_i$ which leads to a linear system:

$$M_{tot}v_{aff} = v_{tot} \tag{3}$$

where M_{tot} and v_{tot} are the data matrix and vector, respectively, and v_{aff} is the vector of the unknown affine parameters (for full details, see [10]).

Given the model of uncertainty of the constituent lines in a structure, the covariances of the output affine parameters can be expressed as follows [12]:

$$\Lambda_{r_s t} = M_{tot}^{-1} \tag{4}$$

where $\Lambda_{r_s t}$ is the 4×4 covariance matrix of the affine parameters r_s and t.

2.1 Tracking Shallow Structures

The affine motion constraint is used in a dynamic model to predict and track shallow structures over time. Tracking requires:

1. A dynamic model of the motion of a structure.

2. A match measure to choose good matches for a structure in every newly acquired frame. The constraints on the search for the potential matches are provided by the dynamic model.

3. A mechanism for fusing the current estimate of the affine motion and the 3D location parameters of a structure with those obtained from the newly acquired data.

The affine motion parameters derived in Equation 3 provide a dynamic model of prediction of the motion of a shallow structure in the image plane. Kalman filtering is used for prediction and recursive estimation, and Mahalanobis distance [7] is used for matching the predictions with potential matches in a newly acquired frame [10].

2.2 Shallow Structure Identification and Reconstruction

In [11] affine tracking is embedded in an algorithm to automatically identify shallow structures. The essential idea is that if a hypothesized structure can be consistently tracked and its 3D depth over time is consistent with a shallow structure model, then the structure is identified as shallow otherwise it is labeled non–shallow. The depth of structures identified as shallow is computed from the scale parameter in the affine transformation of Equation 3. It is represented in the coordinate system of the first frame in the sequence.

3 Pose Determination

Using the depths of the shallow structures recovered by the affine-based algorithm, a partial model of the environment can be built. This model has the same coordinate system as that of the first frame's coordinate system. Given correspondences between model and image tokens in subsequent image frames, the pose parameters (rotation and translation) that relate the subsequent frames' coordinate systems to the model coordinate system can be computed. In an earlier paper [5] least-squares techniques for pose determination were developed. These techniques are optimal with respect to gaussian noise in the input image measurements. In this section, the least-squares techniques are extended to also handle gaussian noise in the 3D model. The techniques presented in this section assume point correspondences but are easily modified for line correspondences.

The rigid body transformation from the world coordinate system to the camera coordinate system can be represented as a rotation (R) followed by a translation ($\vec{T}$). A point $\vec{p}$ in world coordinates gets mapped to the point $\vec{p}_c$ in camera coordinates as:

$$\vec{p}_c = R(\vec{p}) + \vec{T} \tag{5}$$

Using equation (5) and assuming perspective projection, the pose constraint equations for the ith point $\vec{p}_i$ in a set of "m" points can be written in the following manner:

$$\frac{1}{p_{czi}}\vec{C}_{xi} \cdot (R\vec{p}_i + \vec{T}) = 0 \tag{6}$$

$$\frac{1}{p_{czi}}\vec{C}_{yi} \cdot (R\vec{p}_i + \vec{T}) = 0 \tag{7}$$

$$\vec{C}_{xi} = (s_x, 0, -I_{xi}) \tag{8}$$

$$\vec{C}_{yi} = (0, s_y, -I_{yi}) \tag{9}$$

$$p_{czi} = (R\vec{p}_i + \vec{T})_z \tag{10}$$

where (I_{xi}, I_{yi}) is the image projection of the point and (s_x, s_y) is the focal length in pixels along each axis.

The non-linear system of constraint equations for the pose parameters R and $\vec{T}$ is solved using the guass-newton technique [12]. Given a current estimate $R, \vec{T}$, the constraint equations (6,7) are linearized about the estimate:

$$\frac{1}{p_{czi}}(\vec{C}_{xi} \cdot \Delta\vec{T} + \delta\vec{\omega} \cdot \vec{b}_{xi}) = -\frac{1}{p_{czi}}\vec{C}_{xi} \cdot \vec{p}_{ci} + \eta_x \tag{11}$$

$$\frac{1}{p_{czi}}(\vec{C}_{yi} \cdot \Delta\vec{T} + \delta\vec{\omega} \cdot \vec{b}_{yi}) = -\frac{1}{p_{czi}}\vec{C}_{yi} \cdot \vec{p}_{ci} + \eta_y \tag{12}$$

where $\vec{b}_{xi} = R\vec{p}_i \times \vec{C}_{xi}$ and $\vec{b}_{yi} = R\vec{p}_i \times \vec{C}_{yi}$. The above equations relate the pose increments $\delta\omega$ (rotation) and ΔT (translation) to the computed measurement errors using the current pose estimate. The noise terms in the two equations, η_x and η_y are functions of both the 3D model noise $\Delta\vec{p}_i$ and the image noise $\Delta X, \Delta Y$:

$$\eta_x = \Delta X + \frac{1}{p_{czi}}\vec{C}_{xi} \cdot (R(\Delta\vec{p}_i)) \tag{13}$$

$$\eta_y = \Delta Y + \frac{1}{p_{czi}}\vec{C}_{yi} \cdot (R(\Delta\vec{p}_i)) \tag{14}$$

Therefore for the ith point, two such equations (11 and 12) can be written and for a set of "m" points, a total of "$2m$" equations are obtained. At each iteration in the minimization process, the linear system of equations is solved using equation (23) to find the best increment vector[4]. This increment is added to the current pose estimate and the process repeated until there is convergence.

If the correct estimate of pose were known, the measurement noise terms η_x and η_y would be equal to the sum of the measurement error of the image point location and the projection of the error in the model point along the image x-axis and y-axis respectively. The measurements of the image point locations are assumed to be corrupted with identical, independent, zero-mean gaussian noise. The 3D model points are also assumed to be corrupted by zero-mean independent gaussian noise. Thus the covariance matrix "V" corresponding to the noise in the linear system of equations (22) in the Appendix is a band matrix in which the non-zero entries are 2×2 matrices about the diagonal. The output covariance matrix for the pose rotation and translation parameters is given by equation (24) evaluated at the final pose estimate.

4 Induced Stereo

In this section, we present techniques for computing 3D estimates of new points in the world coordinate system from their tracked image locations over a multi-frame sequence. The mathematics for both extending the model and refining the initial modeled points is presented. Computed with the estimate of each new model point is an estimate of the covariance of its error. These covariances are functions of the input image measurement covariances and the initial 3D model point covariances.

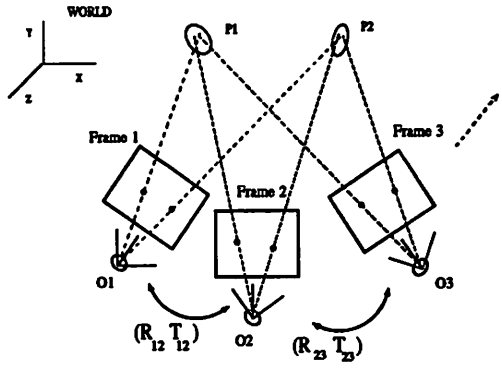

Figure 1: Model Extension and Refinement.

The matching of image features to the partial model is obtained by the tracking method described in Section 2. Given these correspondences, pose estimation is done for each frame using the method presented in the previous

[4]The appendix reviews some salient information on solving over constrained linear equations.

section. Image tokens corresponding to new features are also tracked over a sequence of frames using the computed optic flow between pairs of successive frames [13]. Typically corner points (defined by the intersection of two image lines) are tracked although any image feature which can be reliably tracked may be used. The 3D estimate of the corner point is obtained by the pseudo-intersection of all the image projection rays for a tracked image point. A nice property of the system described here is that the pose estimation process provides the world coordinate frame as a stable coordinate frame in which 3D measurements from a sequence of frames can be combined. Independent measurements can be made relating the coordinate system of each frame in the sequence to the world coordinate frame.

Points are located by the psuedo-intersection process in two steps. In the first step, a 3D error function is minimised to find an initial estimate of the point's location. This step, however, does not yield the optimal estimate since the various error terms are not weighted by the input covariances. In the second step, an image-based error function is optimised in which the error terms are inversely weighted by a combination of the input covariances of the pose estimate and the image measurements.

Let r_i be the unit vector corresponding to the image projection ray for an image point in the *ith* frame. The pose estimation for this frame is given by the rotation R_i and translation $\vec{T}_i$ (see equation (5)). Since the image projection rays do not intersect at a unique point[5], the 3D pseudo-intersection point $\vec{p}$ is obtained by minimising an error function E:

$$E = \sum_{i=1}^{n} \|(R_i(\vec{p}) + \vec{T}_i) \times r_i\|^2 \tag{15}$$

which is the sum of squares of the perpendicular distances from the psuedo-intersection point $\vec{p}$ to the image projection rays. Differentiating E with respect to the unknown variable $\vec{p}$ leads to a set of linear equations, which are then solved to give the initial estimate for $\vec{p}$.

In the second step, the pose constraint equations (6, 7) are used to formulate image-based error equations for the X and Y projections of the model points.

$$\frac{1}{p_{cz}}\vec{C}_{xi} \cdot R_i(\vec{p}) = -\frac{1}{p_{cz}}\vec{C}_{xi} \cdot \vec{T}_i + \zeta_X \tag{16}$$

$$\frac{1}{p_{cz}}\vec{C}_{yi} \cdot R_i(\vec{p}) = -\frac{1}{p_{cz}}\vec{C}_{yi} \cdot \vec{T}_i + \zeta_Y \tag{17}$$

where ζ_X and ζ_Y are the noise terms in the two equations. ζ_X and ζ_Y are functions of both noise in pose $\Delta\vec{T}_i$ and $\delta\vec{\omega}_i$ and image noise $(\Delta X, \Delta Y)$:

$$\zeta_X = \Delta X + \frac{1}{p_{cz}}\vec{C}_{xi} \cdot \Delta\vec{T}_i + \frac{1}{p_{cz}}\delta\vec{\omega}_i \cdot \vec{b}_i \tag{18}$$

$$\zeta_Y = \Delta Y + \frac{1}{p_{cz}}\vec{C}_{yi} \cdot \Delta\vec{T}_i + \frac{1}{p_{cz}}\delta\vec{\omega}_i \cdot \vec{b}_i \tag{19}$$

In this case the 3D model point $\vec{p}$ is the unknown variable. The denominator p_{cz} in the equations (16 and 17) corresponds to the depth of the point and is a function of the unknown variable $\vec{p}$. Therefore, for each frame over which the point is tracked, two non-linear constraint equations (16 and 17) are obtained[6] . An iterative procedure is employed to solve the system of non-linear equations. At each iteration, the denominator p_{cz} is held constant using the previous estimate of $\vec{p}$ and the resulting linear system of equations is solved using equation (23) (see Appendix). The iterative procedure is repeated until there is convergence. In practice, we have found one iteration is sufficient for robust results. The input covariance matrix V required for the normal equations is obtained from the expressions derived above for the noise terms ζ_X, ζ_Y. The output covariance of the 3D point estimate is given by equation (24).

In the batch method, information from all frames is used simultaneously to estimate the 3D locations of tracked image points. However, it may be desired to sequentially update the location of new points after every pair (or a larger set) of frames. In the sequential or quasi-batch mode, equations (16 and 17) are again used to estimate the 3D

[5] Due to noise both in image measurements and pose estimates.

[6] A minimum of two frames is needed to solve the system of equations.

location of image points tracked over the current set of frames. These new estimates must be fused with the previous estimates to obtain the current optimal estimate. The covariance matrices associated with each estimate are used to fuse the two estimates and provide a new uncertainty matrix using the standard Kalman Filtering equations.

Let the estimate of the point's 3D location and its covariance at frame "t_1" be $\vec{p}(t_1)$ and $\Lambda_p(t_1)$ respectively. A new 3D location measurement $\vec{Q}$ with uncertainty (covariance matrix Λ_Q) is computed from a batch of "n" image frames. The fused location estimate $\vec{p}(t_n)$ and updated covariance matrix $\Lambda_p(t_n)$ at frame "t_n" are given by:

$$\vec{p}(t_n) = \Lambda_p(t_n)(\Lambda_p(t_1)^{-1}\vec{p}(t_1) + \Lambda_Q^{-1}\vec{Q}) \quad (20)$$

$$\Lambda_p(t_n) = (\Lambda_p(t_1)^{-1} + \Lambda_Q^{-1})^{-1} \quad (21)$$

This same method is used for model refinement. Initial model points have associated with them their input covariance matrices. When the model is tracked over a new batch of frames, 3D measurements can also be made for the model points by the above psuedo-intersection procedure. These new measurements are fused with the old estimate using the above equation.

5 Experimental Results and Discussion

We now present results over three (real data) multi-frame sequences. In subsection 5.1, results of the model extension and refinement algorithm for the BOX sequence (Figure 2) are presented, and in subsection 5.2, for the sequences A211 and COMP (Figure 5), similar results are presented using an initial model built from a few points on the recovered shallow structures.

The image sequences were captured with a SONY B/W AVC–D1 camera, with an approximate FOV of 24 degrees and digitized to 256–by–242 pixels. In all experiments the image center was assumed to be at the center of the image frame and the effective focal length was calculated from the manufacturers specification sheets. Since we have shown in [6] that errors in the image center do not significantly affect the location of new points in a world coordinate system (for a small field of view imaging system), calibration for the image center has not been done.

5.1 Box Sequence

The BOX sequence, consisting of 8 frames, was generated by rotating the box (Fig. 2) about its central vertical axis in increments of 3.6°, while the camera was kept stationary. The location of 30 points (marked in Fig.2 by circles and crosses) in a world coordinate system was measured to an accuracy of approximately 1 mm along each axis. The depth of the points (in the first frame's coordinate system) used in our experiment varied from 575 mm to 700 mm.

The fifteen points marked by crosses in Figure 2 were used as the initial model to do pose estimation [5] for each frame. Various experiments were performed with different amounts of synthetic uniform noise added to the measured 3D locations of the cross points. Using the computed poses, 3D estimates of the remaining 15 points (marked by circles in Figure 2) were computed. In addition, the initial model of 15 (cross marked) points was refined. The algorithm described in Section 4 was run in a batch mode over all 8 frames to perform these experiments; the results of these experiments are reported in Table 1. The first column of Table 1 gives the range of noise added to the initial model points. Thus a 10 mm entry in the first column means uniform noise in the range of +/- 10 mm was added to each of the 3D coordinates of the model points. The average error[7] of the 15 initial model points for each experiment (prior to any refinement) is given in the second column of Table 1. The third column in the table shows the results of the model refinement process; it gives the average output error of the 15 (now refined) initial model points. The fourth column in the table shows the results of the model extension process; it gives the average output error of the 15 new (circle) points.

As can be seen from the first row in Table 1, the average error for model extension when there is no noise in the initial model is 1.38 mm. The maximum error was 2.6 mm and the minimum error was 0.44 mm. The average

[7]The average error is the root mean square (RMS) value of the 3D location error of all points.

percentage error was 0.25 %. As the noise in the initial model increases, the errors in model extension and refinement also increase. However, except for the first two cases in Table 1, the average output error for both model extension and refinement were significantly lower than the average input error of the initial model points.

The model extension and refinement algorithm was also run in a sequential mode, where new 3D locations were computed after every new pair of frames and the results were fused with previous estimates. Figure 3 shows the results of such an experiment. For this experiment, the range of input noise was 5mm. and the average error of the initial model points was 4.49 mm (corresponding to the fifth row in Table 1). The average output error in location of both the initial model points and the new (circle) 3D points is plotted for every image frame in the sequence. As can be noticed in the figure, the 3D error in both the initial model points and the unknown points monotonically decreases across all frames. The average error of the new points is reduced from 6.5 mm after the first pair of frames to about 3.7 mm at the end. The average error of the initial points is reduced from 4.49 mm to about 2.8 mm.

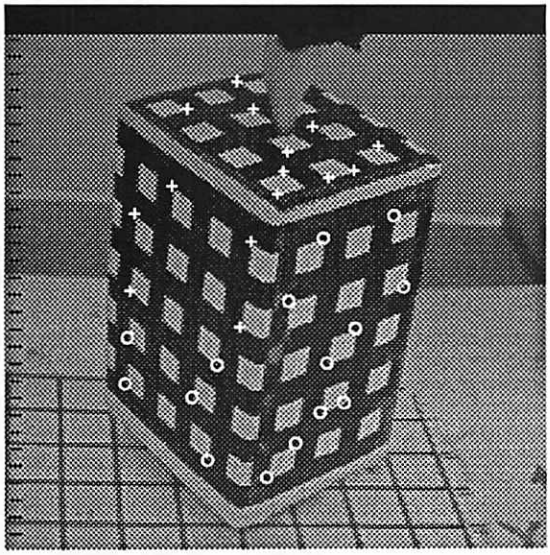

Figure 2: **Box Image.**

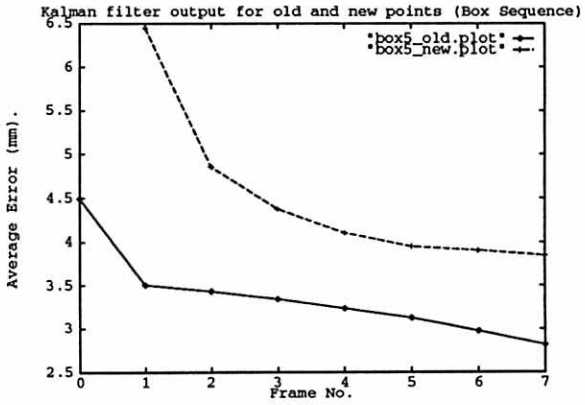

Figure 3: Box Sequence. Plot of average error over the frame sequence for for the new points (Model Extension) and for the initial model points (Model Refinement).

Table 1: **Computed average output 3D location errors for model extension process with noisy input model points for the Box Sequence of 8 frames.** Input Noise to model is synthetic uniform noise.

Range Input Noise	Average Input Noise	Average Output Noise	
		Initial Points	New Points
mm	mm	mm	mm
0	0.00	0.00	1.38
1	1.02	1.01	1.69
2	1.95	1.52	1.92
3	3.06	2.00	2.23
5	4.49	3.00	3.78
7	6.96	3.32	3.84
10	10.25	4.16	6.31
20	17.29	10.32	16.23

5.2 A211 and COMP sequences

The A211 (10 frames) and COMP (6 frames) sequences were generated by taking images from a camera mounted on a mobile robot moving roughly parallel to the optical axis. Figure 5 shows the first frames in the A211 and COMP sequences. Between consecutive frames, the robot was translated approximately 0.38 and 1.4 feet respectively for the A211 and COMP image sequences. The depth of some salient structures in each seqeunce was measured with a tape measure.

Figure 4: **Shallow structures identified in the A211 and COMP image sequences.**

In both sequences, image lines were extracted for each frame using Boldt's [1] line grouping system. The tracking algorithm was applied to the image sequences to identify the shallow structures in the scene. Line triples were automatically selected to hypothesize aggregate structures. Each of these was tested for affine trackability, resulting in its labeling as a shallow or a non–shallow structure [11]. Figure 4 shows in bold lines the structures identified as shallow by the algorithm for the two sequences.

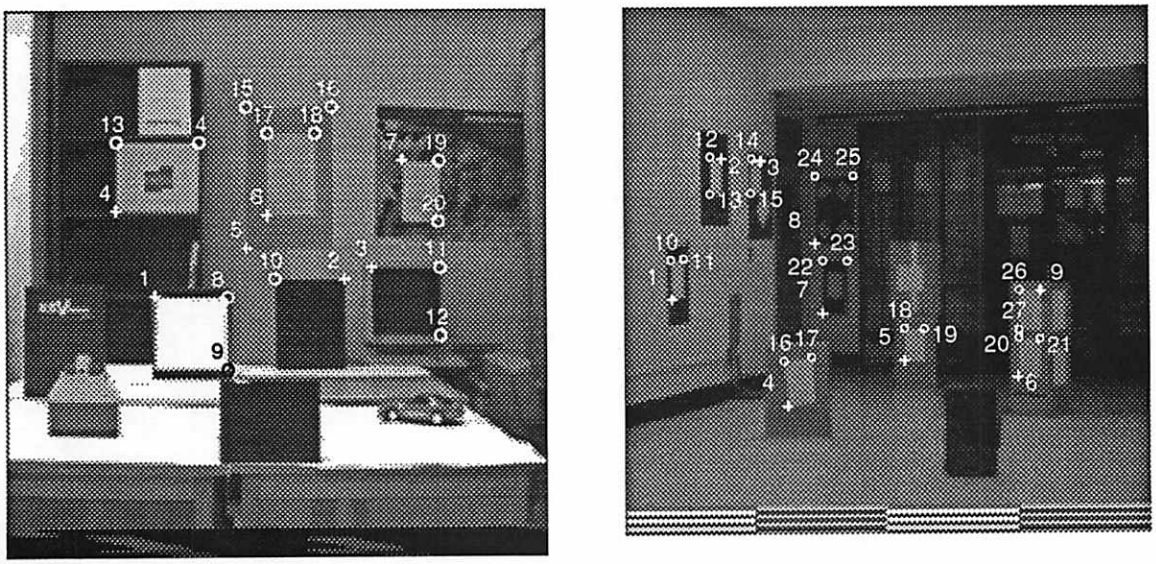

Figure 5: **A211 and COMP Images.**

The numbered points marked by crosses in Figure 5 lying on the recovered shallow structures were used as the initial model points for the A211 and COMP image sequences respectively. These points are defined by the intersection of some of the pairs of lines belonging to shallow structures. The 3D model locations were constructed by back projecting the points in the first image's coordinate frame.

Table 2: **Absolute and Percentage 3D location errors for points in A211 sequence (see Fig. 5.)**

		INPUT		OUTPUT	
Pt. No.	Depth ft.	Abs. Err. ft.	% Err.	Abs. Err. ft.	% Err.
Initial Points					
1	13.4	0.24	1.80 %	0.24	1.78 %
2	14.6	0.19	1.31 %	0.20	1.34 %
3	19.0	0.74	3.88 %	0.66	3.46 %
4	19.0	0.16	0.86 %	0.11	0.60 %
5	20.4	0.13	0.62 %	0.17	0.86 %
6	20.4	0.39	1.90 %	0.32	1.60 %
7	20.4	0.49	2.38 %	0.46	2.25 %
New Points					
8	13.4	-	-	0.11	0.79 %
9	13.4	-	-	0.00	0.01 %
10	14.6	-	-	0.53	3.65 %
11	19.0	-	-	0.73	3.86 %
12	19.0	-	-	0.54	2.82 %
13	19.0	-	-	0.11	0.59 %
14	19.0	-	-	0.07	0.34 %
15	20.4	-	-	0.23	1.13 %
16	20.4	-	-	0.27	1.32 %
17	20.4	-	-	0.12	0.57 %
18	20.4	-	-	0.34	1.65 %
19	20.4	-	-	0.62	3.02 %
20	20.4	-	-	0.59	2.92 %

The model extension and refinement algorithm was run in a sequential mode. Tables 2 and 3 show the result of locating new points (circled and numbered in Figure 5) and refining the initial model points (marked by crosses in the figures). The ground truth available for both experiments was only the depths (as opposed to 3D location) of the points in the first image's coordinate frame. Thus the results shown in Tables 2 and 3 compare the measured depth value (ground truth) with the recovered depth value. Column 2 in the tables shows the measured depth of the point in the first image coordinate frame. Columns 3 and 4 show the output error and percentage error in depth, respectively, for the initial model points as recovered by the affine-based tracking algorithm. Columns 5 and 6 show the output error and percentage error in depth (after model refinement and extension) respectively. For the new points, it is assumed that no initial model was available, therefore columns 3 and 4 for these points are blank. Note that these points also belong to the reconstructed shallow structures. However, their reconstructed locations were not used as a part of the initial partial model. Instead, these points were used to demonstrate model extension because the ground truth was available only for these structures. In the tables, the percentage error in depth is computed with respect to the depth in the first image's coordinate frame.

The average input error in depths of the seven initial model points in the A211 sequence (as recovered by the affine-based tracking algorithm) was 0.4 feet (1.85 % error). At the end of the ten frames, the average error of the 7 initial points was 0.37 feet (1.76 %). The thirteen new points were located to an average accuracy of 0.4 feet (1.63 %).

Table 3: **Absolute and Percentage 3D location errors for points in COMP sequence (see Fig. 5.)**

		INPUT		OUTPUT	
Pt. No.	Depth ft.	Abs. Err. ft.	% Err.	Abs. Err. ft.	% Err.
Initial Points					
1	29.3	-0.23	0.80 %	-0.11	0.36 %
2	31.3	0.26	0.84 %	0.17	0.55 %
3	34.2	-0.10	0.29 %	-0.07	0.20 %
4	25.7	-0.26	1.03 %	-0.23	0.88 %
5	35.8	1.59	4.43 %	1.54	4.31 %
6	28.7	0.39	1.36 %	0.39	1.35 %
7	43.2	-1.65	3.82 %	-1.63	3.76 %
8	43.2	1.18	2.73 %	1.15	2.66 %
9	28.7	1.46	5.08 %	1.41	4.91 %
New Points					
10	29.3	-	-	0.25	0.86 %
11	29.3	-	-	-0.35	1.19 %
12	31.3	-	-	0.51	1.63 %
13	31.3	-	-	0.28	0.89 %
14	34.2	-	-	0.93	2.70 %
15	34.2	-	-	1.31	3.82 %
16	25.7	-	-	-0.02	0.07 %
17	25.7	-	-	0.03	0.11 %
18	35.8	-	-	1.05	2.93 %
19	35.8	-	-	0.50	1.40 %
20	28.7	-	-	-0.11	0.39 %
21	28.7	-	-	0.08	0.29 %
22	43.2	-	-	0.46	1.07 %
23	43.2	-	-	1.77	4.10 %
24	43.2	-	-	-0.45	1.04 %
25	43.2	-	-	0.13	0.30 %
26	28.7	-	-	0.80	2.77 %
27	28.7	-	-	0.25	0.88 %

The average input error in depths of the nine initial model points in the COMP sequence (as recovered by the affine-based tracking algorithm) was 1.01 feet (2.27 % error). At the end of the six frames, the average error of the 9 initial points was 0.98 feet (2.11 %). The eighteen new points were located to an average accuracy of 0.69 feet (1.46 %). For this experiment the measured depth values are only approximate to about 0.5 feet for some points. This is especially true for points lying on the left side wall (points 1, 2, 3 etc. in Figure 5).

In both experiments, the model extension process was fairly accurate in locating new points. However, there was only slight improvement in the initial model as a result of the model refinement process. The robust recovery of the location of new 3D points depends on the camera motion. Optimal angles for triangulation are achieved when there is significant translation parallel to the image plane. For this reason, the best results obtained by the model extension and refinement process were for the BOX sequence. In the A211 and COMP sequence, the translation of the camera is mostly along the optical axis. Thus, the FOE (focus of expansion) lies on the image plane. Points close to the FOE have hardly any disparity and their depths cannot be reliably estimated.

Finally, the accuracy of the model extension process depends on the initial accuracy of the model points. If the initial model points have a large amount of noise, then the poses determined for any batch of frames will be highly

correlated. In this case, the 3D location estimates of new points will be correlated both across all points and also all frames. To fully account for this correlation, covariance matrices equal to the size of number of points times number of frames will have to be inverted. In our case, it is assumed that the initial points do not have significant noise and hence the cross-correlations can be ignored. But for larger amounts of noise, it may not be possible to ignore these effects [8].

Appendix

Some facts from linear system estimation theory are reviewed. An unknown parameter vector $\vec{x}$ with "p" elements is related to a set of "n" noisy observations $\vec{y}$ by the following equation:

$$A\vec{x} = \vec{y} + \vec{\eta} \tag{22}$$

where $\vec{\eta}$ is zero-mean Gaussian noise with covariance matrix V. Assume, that this set of equations is an over-constrained system. Then the Best Linear Unbiased Estimate (BLUE) [12] of the unknown vector $\vec{x}$ and the covariance matrix "P" of the output parameters are given by:

$$\hat{x} = (A^T V^{-1} A)^{-1} A^T V^{-1} y \tag{23}$$

$$P = (A^T V^{-1} A)^{-1} \tag{24}$$

References

[1] M. Boldt and R.Weiss and E. Riseman, "Token-based Extraction of Straight Lines," *IEEE Transactions on Systems Man and Cybernetics*, volume 19, no. 6, pp. 1581–1594, 1989.

[2] T. J. Broida and R. Chellappa, "Estimating the Kinematics and Structure of a Rigid Object from a Sequence of Monocular Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 6, pp. 497–513, 1991.

[3] S. Chandraskekhar and R. Chellapa, "A Two-Step Approach to Passive Navigation Using a Monocular Image Sequence," *USC-SIPI Technical Report 170*, University of Southern California, Electrical Engineering-Systems, 1991.

[4] N. Cui, J. Weng and P. Cohen, "Extended structure and motion analysis from monocular image sequences," *Proceedings 3rd IEEE International Conference on Computer Vision*, Osaka, Japan, Dec. 1990.

[5] R. Kumar and A.R. Hanson, "Robust Estimation of Camera Location and Orientation from Noisy Data with Outliers," *Proc. IEEE Workshop on Interpretation of 3D scenes*, Austin, Texas, Nov. 1989.

[6] R. Kumar and A.R. Hanson, "Sensitivity of pose refinement to accurate estimation of camera parameters," *IEEE International Conference on Computer Vision*, Osaka, Japan, Dec. 1990.

[7] P. C. Mahalanobis, "On the Generalised Distance in Statistics," *Proceedings of the National Institute of Science, India*, (12), pp. 49–55, 1936.

[8] J. Oliensis and J. I. Thomas, "Incorporating motion error in multi-frame structure from motion", *Proceedings IEEE Workshop on Visual Motion*, Princeton, N.J., Oct. 1991.

[9] H. S. Sawhney and A. R. Hanson, "Comparative Results of Some Motion Algorithms on Real Image Sequences," *Proc. DARPA Image Understanding Workshop*, 1990.

[10] H. S. Sawhney and A. R. Hanson, "Identification and 3D description of 'shallow' environmental structure in a sequence of images", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 179–186, Hawaii, June 1991.

[11] H. S. Sawhney and A. R. Hanson, "Affine Trackability aids Obstacle Detection", in this proceedings.

[12] Gilbert Strang, "Introduction to Applied Mathematics," Wellesley–Cambridge Press, MA, 1986.

[13] L. R. Williams and A. R. Hanson, "Translating Optical Flow into Token Matches and Depth from Looming", *Second Int. Conf. on Computer Vision*, pp. 441-448, 1989.