

An Integrated Approach for Segmentation and Estimation of Planar Structures

Jonathan Alon and Stan Sclaroff
 Computer Science Department
 Boston University
 Boston, MA 02215

Abstract

Standard structure from motion algorithms recover 3D structure of points. If a surface representation is desired, for example a piece-wise planar representation, then a two-step procedure typically follows: in the first step the plane-membership of points is first determined manually, and in a subsequent step planes are fitted to the sets of points thus determined, and their parameters are recovered. This paper presents an approach for automatically segmenting planar structures from a sequence of images, and simultaneously estimating their parameters. In the proposed approach the plane-membership of points is determined automatically, and the planar structure parameters are recovered directly in the algorithm rather than indirectly in a post-processing stage. Simulated and real experimental results show the efficacy of this approach.

1 Introduction

This paper addresses the problem of segmenting multiple planar structures from an image sequence, and estimating their model parameters. It is assumed that the images are taken from different view points by a perspective camera, and that the planar structures contain regions of sufficient intensity variation, so that point features can be extracted and matched reliably across frames. Planar surfaces are important because they are common in both indoor and outdoor environments. Planar man-made structures such as floors, walls, buildings and sidewalks occur frequently in real image sequences.

A planar surface has three degrees of freedom, and can therefore be represented with three parameters, for example two parameters for its unit normal and one parameter for its distance from the origin. (This is not the only possible representation.) Given a set of three-dimensional points, these parameters can easily be recovered using least-squares regression methods. In fact, these methods are widely applied as a post-processing phase in many structure from motion algorithms; first the 3D structure of points is recovered, and then a plane is fit to this set of points. However, when a scene consists of multiple planes, the correct assignment of points to planes has to be determined first. This procedure is typically performed manually. In the proposed approach, these parameters are recovered directly in the algorithm rather than indirectly in a post-processing phase.

The proposed approach builds upon previous work by Darrell [5]. While Darrell's main motivation is motion segmentation, the purpose of the work presented here is segmentation and estimation of planar structures. In this approach multiple planar hypotheses are initially formed by generating random groups of points or support maps. The parameters of every planar hypothesis are estimated with a Kalman filter, and the support maps are then updated by thresholding the reprojection error between the observed and resynthesized feature tracks. In the final step of the procedure, a subset of hypotheses that best accounts for the observed tracks is selected and the estimates are refined. In this paper only segmentation of multiple planar surfaces is considered, but the approach described is more general and can handle other surfaces that can be parameterized as $z = f(x, y)$. Also, other planar structure recovery algorithms can be incorporated in the same formulation.

2 Related Work

It is well known that the mapping between two projected views of a plane is completely specified by a 3x3 matrix [15]. The group of these matrices is called the planar projective group. Every member of this group has eight degrees of freedom [20], and can therefore be determined with four point correspondences. Once these eight parameters are computed, the actual structure and motion parameters can be estimated by computing the singular value decomposition of the 3x3 matrix [21]. The number of solutions, which is one or two, depends on the multiplicity of the singular values. A unique recovery of structure from motion can be obtained via correspondences of four points on a plane and two points not on that plane [10]. Improved quality of planar structure estimates can be obtained by including lines, texture and even hallucinated point correspondences in the formulation ([17], [7], [18]).

Segmentation of planar regions has recently been applied in various situations using different geometric constraints. These include the detection of planar regions using projective invariants [16], the estimation of vanishing points and lines using commonly occurring types of geometric groupings, such as equally spaced coplanar parallel lines [14], the matching of facets in pairs of images using chains of corners [19], and the reconstruction of piecewise planar models using a single 3D line with a textured neighborhood [2]. The application of the work presented in this

paper is aimed at but not restricted to the reconstruction of piecewise planar models. The geometric constraints used are point features correspondences.

3 Planar Structure Recovery

In this section, our recursive estimation framework to recover planar structure and camera motion is briefly restated from [1]. The measurement vector of the Extended Kalman Filter (EKF) is given by

$$\mathbf{z} = (u_1, v_1, u_2, v_2, \dots, u_N, v_N), \quad (1)$$

where (u_i, v_i) is the image location of the i^{th} feature point, and $i \in \{1 \dots N\}$ where N is the number of features. The state vector of the EKF is given by

$$\mathbf{x} = (t_X, t_Y, t_Z, \beta, \omega_X, \omega_Y, \omega_Z, \theta, \phi), \quad (2)$$

where (t_X, t_Y, t_Z) is the translation vector, $(\omega_X, \omega_Y, \omega_Z)$ is the incremental rotation vector, β is the inverse focal length, and θ and ϕ are the tilt and slant angles of the plane. The plane unit normal is represented with these angles:

$$\begin{aligned} n_X &= \sin(\theta) \cos(\phi) \\ n_Y &= \sin(\theta) \sin(\phi) \\ n_Z &= \cos(\theta) \cos(\phi) \end{aligned} \quad (3)$$

The dynamic model in the EKF is chosen trivially as the identity plus noise:

$$\mathbf{x}(k+1) = \mathbf{x}(k) + v; \quad (4)$$

where the process noise v is a zero mean Gaussian random vector with covariance matrix Q . The measurement model is given by

$$\mathbf{z}(k) = m(\mathbf{x}(k), \mathbf{z}(1)) + \omega \quad (5)$$

where the measurement noise ω is a zero mean Gaussian random vector with covariance matrix R . The derivation of the non-linear measurement function m , and the associated measurement Jacobian H are given in [1].

The above state \mathbf{x} consists of nine parameters; however if the inverse focal length is excluded from the estimation, then the state has only eight degrees of freedom, which can be determined with four¹ point correspondences. Also, it is well known that in a monocular image sequence taken by a perspective camera, structure can only be recovered up to a scale factor. In the context of planar structure recovery, this means that the distance of the plane from the origin cannot be determined, and is therefore fixed for the purpose of gaining a solution.

¹It is assumed that no three of the four points are collinear.

4 Planar Structure Segmentation

The previous section described how to recover the parameters of a single plane in a scene. This section shows how multiple planes can be automatically segmented, without any prior knowledge about the number of planes in the scene or their orientations. In addition, the true correspondence between point features and planes is assumed unknown. The approach described here builds upon the approach previously proposed by Darrell [5]. This approach consists of two steps: multiple hypothesis testing and hypothesis selection, which are described next. The main difference is that in this paper segmentation of multiple planar structures is computed rather than segmentation of multiple independent motions. Also, an additional processing phase reduces the number of hypotheses considered before the hypothesis selection stage. This reduces the time required for finding the number of planes in the scene.

4.1 Multiple Hypothesis Testing

The first step greatly resembles the RANSAC procedure [8]; rather than using as much of the data as possible to obtain an initial estimate and then attempting to eliminate the invalid data points, RANSAC uses as small an initial data set as feasible and enlarge this set with consistent data when possible. Initial hypotheses are generated by taking random samples of sets of features, and using those to compute structure estimates. To construct a particular set of features, a feature is first selected at random and is included in the set. The nearest feature to the set is found, and with probability p is included in it. The next nearest feature is then selected, and the process repeats until n features have been selected. The control parameter p is set according to knowledge about the proximity of features that come from the same planes in the scene. The parameter n has to be greater than or equal to the number of features required to constrain the problem. In our case $n \geq 4$ if the focal length is assumed known, or $n \geq 5$ if the focal length is unknown. The set of hypotheses thus generated can be described mathematically by a binary support map

$$s_{ij} = \begin{cases} 1 & \text{if } f_j \in h_i \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where f_j is feature j and h_i is hypothesis i .

Given a particular set of features, an estimate of structure and motion needs to be computed based on only those features in the set as described in Sec. 3. This can be obtained by setting the measurement noise of the features not included in the set to an infinitely large value. This will result in numerical instabilities. As a practical matter it is common to use the information form of the Kalman filter in such instances. The information filter has the additional benefit that if the dimension of the measurement vector is larger than the dimension of the state vector the filter runs

in time quadratic in the number of features, rather than in cubic time using the standard Kalman filter.

After estimating motion and structure parameters for a hypothesized plane based on the initial support, one can compute an updated support that indicates which feature tracks including those not in the initial support, can be considered to be moving rigidly on that plane with the same motion parameters. The new support is computed by thresholding a distance function, as follows:

$$s_{ij} = \begin{cases} 1 & \text{if } D(h_i, f_j) \leq \theta \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where $D(h_i, f_j)$ is the mean of the squared differences between the predicted feature locations for feature j given hypothesis i and the feature locations actually observed:

$$D(h_i, f_j) = \frac{1}{N - N_0 + 1} \sum_{k=N_0}^N (z_j(k) - \hat{z}_j^{(i)}(k))^2 \quad (8)$$

where N is the number of frames, N_0 is the first frame after which the filter starts to converge, $z_j(k)$ is the observed location of feature j in frame k , and $\hat{z}_j^{(i)}(k)$ is the predicted location of feature j given hypothesis i in frame k . the predicted location $\hat{z}_j^{(i)}(k)$ is computed by plugging in the location of feature j in frame 1 and the estimate at frame k into the Kalman measurement equation m :

$$\hat{z}_j^{(i)}(k) = m(z_j(1), \hat{x}^{(i)}(k)) \quad (9)$$

For each hypothesis i a separate Kalman filter is run, and then for each feature j all the differences (also known as innovations in the context of estimation theory) are computed on the fly. The threshold θ is set to a few pixels. Setting θ too low will result in hypotheses with little support and few outliers, while setting it high will result in hypotheses with large support but more outliers. As a post-processing step, the number of hypotheses in the set is reduced by simply discarding all the hypotheses that have too little support; more specifically, hypotheses that include less than n features are discarded. Applying this post-processing step reduces the computational time of the hypothesis selection procedure, which is described next.

4.2 Hypothesis Selection

We now define the objective function for selecting the hypotheses that best explain the data. This function can be derived in a more general form using Bayesian estimation theory and information theory. The reader is referred to [5] for this derivation. In short, the more general derivation involves prior probabilities of the different hypotheses and the observed data, and their respective encoding savings in a information theoretic sense. Since such probabilities are not available, we prefer to approach the derivation from a deterministic optimization viewpoint.

The selection problem can be stated as follows: given a set H of M hypothesized planar structures:

$$H = \{h_0, h_1, \dots, h_M\}. \quad (10)$$

Each h_i contains motion and planar structure estimates. We try to select the subset $\mathcal{L} \subset H$ that best ‘‘covers’’ all the feature tracks using the smallest number of hypotheses.

Skipping the complete derivation, we define the objective function to be maximized:

$$E(\mathbf{a}) = \sum_i \sigma(a_i) \sum_j \left[s_{ij} - \sum_{k \neq i} \sigma(a_k) s_{kj} \right]_{>0} - O \sum_i \sigma(a_i) \quad (11)$$

This function encourages hypotheses which support many features, subject to an inhibitory term that prevents more than one hypothesis from covering the same data, and a term that discourages hypotheses with too little support (e.g. less than O). The selection vector \mathbf{a} that is being estimated represents the subset \mathcal{L} . The sign of each a_i indicates whether the hypothesis $h_i \in \mathcal{L}$: a positive value indicates $h_i \in \mathcal{L}$, and a negative value indicates $h_i \notin \mathcal{L}$. The function $[\cdot]_{>0}$ is defined as follows:

$$[x]_{>0} = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

The function $\sigma(\cdot)$ transforms each a_i into a weight between 0 and 1. Ideally, σ would be the unit step function centered at the origin. However this hard non-linearity makes it very difficult to maximize $E(\mathbf{a})$. Instead the ‘‘softer’’ sigmoid function is used,

$$\sigma(x; \alpha) = \frac{1}{1 + e^{-\alpha x}} \quad (13)$$

where α determines the hardness of the non-linearity: in the limit when $\alpha \rightarrow \infty$, $\lim_{\alpha \rightarrow \infty} \sigma(x; \alpha) = u(x)$, where $u(x)$ is the unit step function. In our implementation we fix $\alpha = 1$, but if the optimization procedure gets trapped frequently in local maxima, then α can be changed gradually in a continuation method. Small values of α result in a smooth cost function for which a rough estimate of the optimal solution can be computed. This solution, which has a better chance of being closer to the global maximum than any other arbitrary guess, serves as the initial estimate for subsequent iterations. Increasing the value of α in subsequent iterations will result in the emergence of the finer details of the cost function, with the hope that a more accurate estimate will eventually be obtained.

A straight forward gradient ascent technique is suitable for finding good solutions to the optimization problem (Eq. 11). The authors in [5] made use of a forward Euler discretization procedure. In our implementation we make use of Powell’s method as described in [22]. Powell’s method is a multidimensional optimization technique

based on a sequence of line optimizations. Although Powell’s method does not make use of gradient information, we found that for a cost of perhaps more computational time, the local maxima found using this technique corresponded to correct estimates of the number of planes in the scene, and to correct estimates of their corresponding supports.

5 Experiments

We now describe two performance experiments. The first experiment is run on synthetic data and its purpose is to test the robustness of the system to measurement noise. The second experiment is run on a real sequence of a box. Both the motion of the camera and the structure are similar in both experiments so the results depicted in the graphs can be qualitatively compared in the absence of ground truth for the real sequence.

5.1 Experiment 1 : Increasing Noise Level

In this experiment we test the robustness of the system to measurement noise for the case where there are multiple planes in the scene. The scene consists of three mutually perpendicular planar structures with checkered patterns overlaid. A total of forty eight features (sixteen on every plane) are used as measurements. The features are placed on the checkered rectilinear grid with uniform spacing. The camera is moving along a circular arc with center at the intersection of the three planes. The radius of the arc is 4 units, and the total angle of rotation is 50° . The camera’s height with respect to the bottom plane is 2 units. The right plane’s unit normal is $(0.342, -0.420, 0.840)$, the left plane’s unit normal is $(-0.940, -0.153, 0.306)$, and the bottom plane’s normal is $(0, 0.894, 0.447)$. The camera’s field of view is 50° which corresponds to $\beta = 0.466$. In each trial, uniform noise with varying standard deviation is added to both u and v image coordinates. The standard deviation corresponds to 0, 2, and 4 pixels, based on an image size of $(512, 512)$.

Example frames are shown in Fig. 1. Thirty hypotheses were randomly generated with the following parameters: the number of features per hypothesis is $n = 6$, the probability of including the next closest feature in the hypothesis is $p = 0.5$. Eight of the initial hypotheses are depicted in Fig. 2. Updated supports were computed using a threshold of 4 pixels based on a window size of $(512, 512)$. Final supports were computed with a threshold of 8 pixels and an MDL penalty term $O = 5$. The three final supports that were computed for this example are shown in Fig. 3.

Graphs of recovered planar structure and camera motion for the three noise levels is shown in Figs. 4, 5, and 6. Hundred random trials were conducted at each noise level, and the average estimates were plotted on the graphs. Table 1 gives statistics for the experiment. As can be seen in both the graphs and the table, the system is robust to noise level increase. In this experiment the system detected the

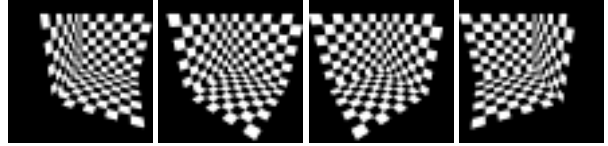


Figure 1: Four of fifty synthetically generated frames of three planes with a textured checker pattern.

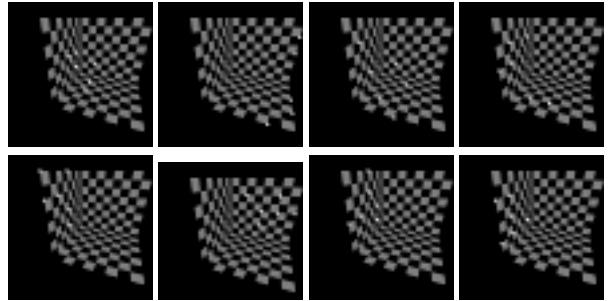


Figure 2: Eight of thirty support maps for planar structure hypotheses. Forty eight features (sixteen from each plane) are tracked over the sequence. Top row : hypotheses that will degenerate. Bottom row: hypotheses that will be validated.

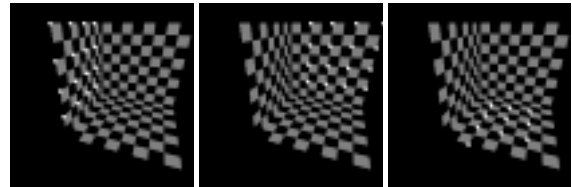


Figure 3: The 3 final support maps for planar structure hypotheses.

		Performance Statistics				
Noise pixels	Plane	Motion Estimation Error				Convergence r_s
		m_t	σ_t	m_q	σ_q	
0	Right	-0.0002	0.0594	-0.0058	0.0206	9
	Left	0.0412	0.1625	-0.0655	0.0949	31
	Bottom	0.0451	0.1404	-0.0175	0.0242	39
2	Right	0.0029	0.0834	-0.0050	0.0261	11
	Left	0.0165	0.0904	-0.0179	0.0426	19
	Bottom	0.0466	0.2063	-0.0139	0.0956	19
4	Right	0.0020	0.0825	-0.0055	0.0258	19
	Left	0.0240	0.1917	-0.0215	0.0572	23
	Bottom	0.0380	0.1311	-0.0081	0.0243	22

Table 1: Average performance statistics for synthetic data experiments with increasing noise level. Experiments were conducted in trials with varying uniform noise (standard deviation 0, 2, and 4 pixels). Mean error and root mean squared error are shown for the recovered camera motion parameters (translation, rotation). For the static parameters (structure) the table provides the frame number for which the normal converges to within 1° of its true value.

correct number of planes in the scene 100/88/85 percent of the time when 0/2/4 pixels uniform noise was added. Structure estimates converged to within 1° of ground truth after 10-30 frames on average. Motion (RMS) errors depicted in Table 1 are perhaps less indicative compared to the results of the single plane case, because here they are dependent on the particular initial estimates (out of nine possible guesses) that caused the filter to converge.

5.2 Experiment 2: Real Sequence

In this section we show an example of a real sequence experiment with the proposed system. The sequence was taken with an off-the-shelf NTSC video. The parameters were set to the same values that were set for the synthetic experiment. Fig. 7 shows example frames from the sequence. Sixteen corners (eight from top plane and eight from front plane) were hand picked and tracked through the entire sequence. All sixteen features were correctly segmented. Graphs showing the translation, rotation and structure of the two segmented planes are shown in Fig. 8. As mentioned above, since ground truth could not be obtained for this sequence, the camera motion and structure of this sequence were designed to be similar to those of the synthetic experiments so the results depicted in the graphs can be compared qualitatively. Indeed, the translation and rotation graphs appear to be very similar. The translation of the front and the top plane differ by a scale factor. This is to be expected because only the direction of the translation can be obtained from a monocular sequence. The normals of the planes appear to converge to the true estimates.

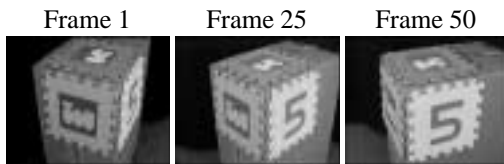


Figure 7: Three of fifty frames of a box sequence. Sixteen features (eight on top plane and eight on front plane) were tracked over the sequence.

6 Conclusion

This paper presented an approach for segmenting a scene that consists of multiple planar structures. The main contribution of our work is that we have shown that the orientations of planar structures can be computed directly without the need to employ a widely used post-processing stage in which planes are fitted to manually segmented sets of three-dimensional points. If good estimates of structure and motion are recovered then the proposed algorithm will typically segment out the correct number of planes in the scene. Good initial guesses for the planar structure orientation are also required, but these can be obtained by running

a batch algorithm on the first few frames of the sequence.

In order to have a complete end-to-end system, that is capable of reconstructing three-dimensional scenes from image sequences, there are few missing parts that have to be added to the proposed system. On the one end there is a need to incorporate structure and motion estimates to guide the feature matching process in a manner similar to [3]. Handling missing features and incorporating new features in the formulation is equally important, and a few promising attempts [4, 12] in this direction have been made. On the other end, there is a need to merge all the structure estimates into a meaningful textured 3D model. Recent work in the areas of 3D model reconstruction [6, 11, 9] and image-consistent surface triangulation [13] seem to be promising.

References

- [1] The Authors. In *CVPR*, II:550–556, 2000.
- [2] C. Baillard and A. Zisserman. Automatic reconstruction of piecewise planar models from multiple views. In *CVPR*, II:559–565, 1999.
- [3] P.A. Beardsley, A. Zisserman, and D.W. Murray. Sequential updating of projective and affine structure from motion. *IJCV*, 23(3):235–259, 1997.
- [4] A. Chiuso and S. Soatto. 3D motion and structure from 2D motion causally integrated over time : Analysis. In *IEEE Trans. Robotics and Automation*, 2000.
- [5] T. Darrell and A.P. Pentland. Cooperative robust estimation using layers of support. *PAMI*, 17(5):474–487, May 1995.
- [6] P.E. Debevec, C.J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *SIGGRAPH*, 1996.
- [7] C. Thorpe F. Dellaert and S. Thrun. Super-resolved texture tracking of planar surface patches. In *IEEE/RSJ Conf. on Intelligent Robotic Systems*, 1998.
- [8] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *CACM*, 24(6):381–395, 1981.
- [9] M. Han and T. Kanade. Creating 3D models with uncalibrated cameras. In *WACV*, 2000.
- [10] C.H. Lee. Structure and motion from two perspective views via planar patch. In *ICCV*, 158–164, 1988.
- [11] A. Manassis, A. Hilton, P. Palmer, P. McLauchlan, and X. Shen. Reconstruction of scene models from sparse 3D structure. In *CVPR*, II:666–671, 2000.
- [12] P.F. McLauchlan. The variable state dimension filter. Tech. Report VSSP 4/99, U. Surrey, Dept. of E.E., 1999.
- [13] D.D. Morris and T. Kanade. Image-consistent surface triangulation. In *CVPR*, I:332–338, 2000.
- [14] F. Schaffalitzky and A. Zisserman. Planar grouping for automatic detection of vanishing lines and points. *IVC*, 18(9):647–658, 2000.

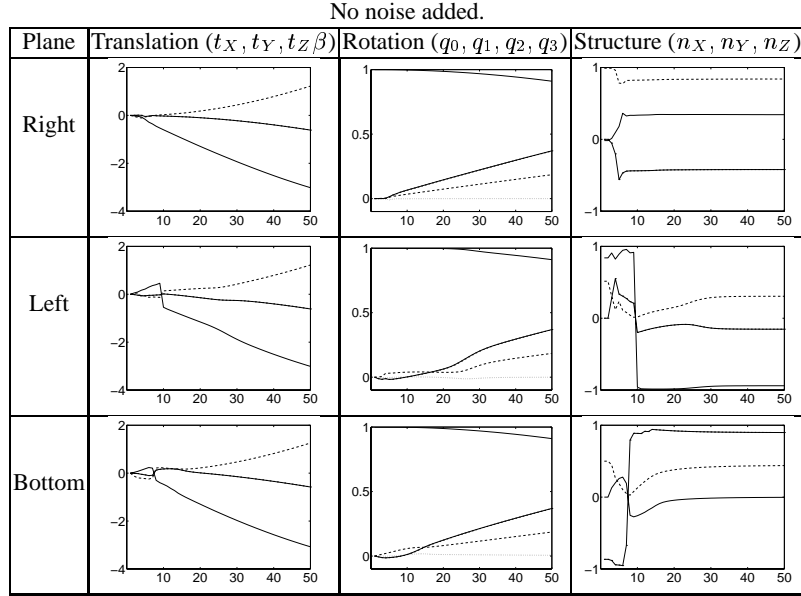


Figure 4: Experiment using synthetic data with no noise added. Multiple trials (hundred at each noise levels) were conducted, and the average estimates for the three planes are shown in the graphs. Each graph's x-axis is the frame number and the y-axis is the state variable. Parameters t_X, q_0, n_X are represented by solid lines on the graphs; t_Y, q_2, n_Y are represented by dash-dot lines; t_Z, q_3, n_Z are represented by dashed lines; and q_1 is represented by a dotted line. For a summary of statistics, see Table 1.

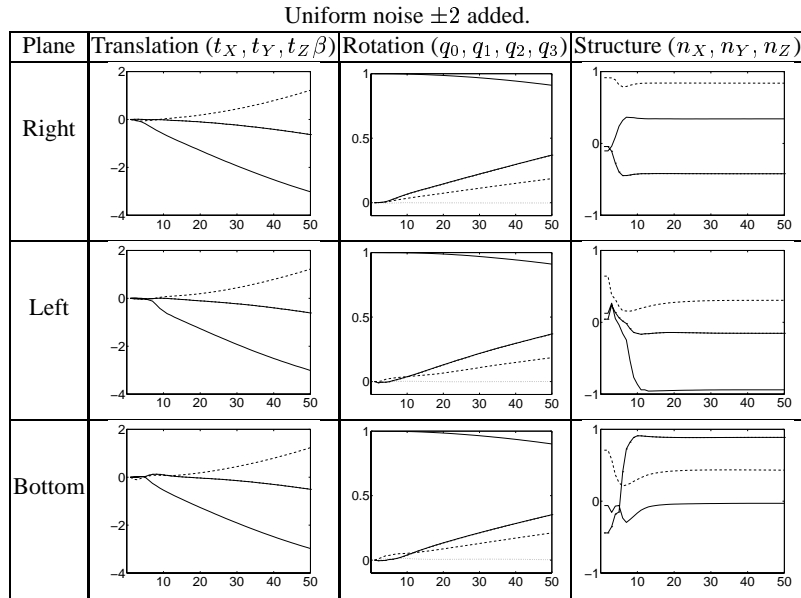


Figure 5: Experiment using synthetic data with 2 pixels uniform noise added.

Uniform noise ± 4 added.

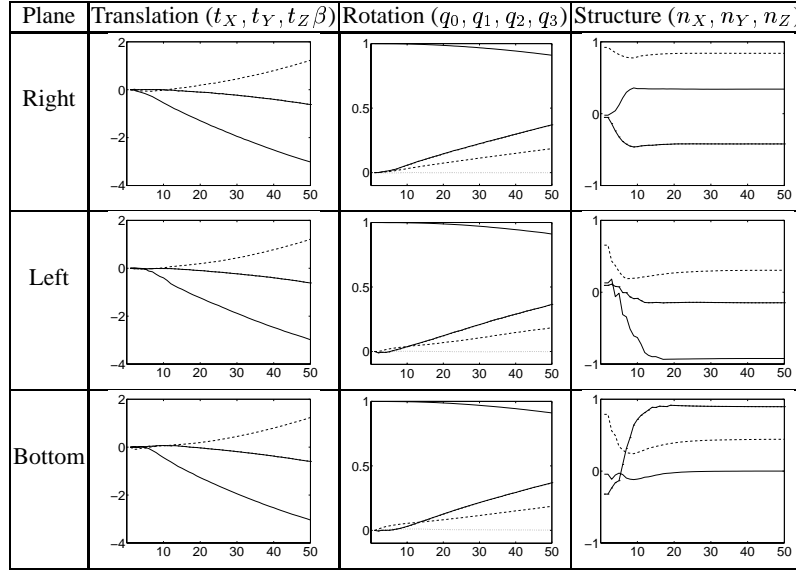


Figure 6: Experiment using synthetic data with 4 pixels uniform noise added.

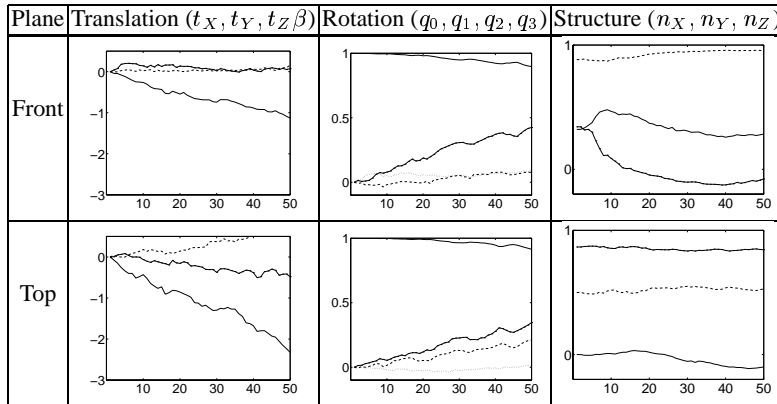


Figure 8: Experiment using box sequence. The estimates for the front and top planes are shown in the graphs. Each graph's x-axis is the frame number and the y-axis is the state variable. Parameters t_X, q_0, n_X are represented by solid lines on the graphs; t_Y, q_2, n_Y are represented by dash-dot lines; t_Z, β, q_3, n_Z are represented by dashed lines; and q_1 is represented by a dotted line.

- [15] J.G. Semple and G.T. Kneebone. *Algebraic Projective Geometry*. Oxford University Press, 1952.
- [16] D. Sinclair and A. Blake. Quantitative planar region detection. *IJCV*, 18(1):77–91, 1996.
- [17] M.E. Spetsakis and Y. Aloimonos. Closed form solution to the structure from motion problem from line correspondences. In *AAAI*, 738–743, 1990.
- [18] R. Szeliski and P.H.S. Torr. Geometrically constrained structure from motion: Points on planes. In *SMILE*, 1998.
- [19] L. Theiler and H. Chabbi. Facet matching from an uncalibrated pair of images. In *MVA*, 1998.
- [20] R.Y. Tsai and T.S. Huang. Estimating 3D motion parameters of a rigid planar patch I. *ASSP*, 29(12):1147–1152, 1981.
- [21] R.Y. Tsai, T.S. Huang, and W.L. Zhu. Estimating 3D motion parameters of a rigid planar patch II: Singular value decomposition. *ASSP*, 30(8):525–533, 1982.
- [22] S. Teukolsky W. Press, B. Flannery and W. Vetterling. *Numerical Recipes in C*. Cambridge University Press, Cambridge, UK, 1988.