

REDUCTION OF INHERENT AMBIGUITIES IN STRUCTURE FROM MOTION PROBLEM USING INERTIAL DATA

G. Qian Q. Zheng and R. Chellappa

Center for Automation Research
Department of Electrical and Computer Engineering
University of Maryland
College Park, MD 20742-3275
{gqian,kale,rama}@cfar.umd.edu

ABSTRACT

In this paper, the reduction of inherent ambiguities in Structure from Motion (SfM) using inertial data is addressed. First, we show that the translation-rotation ambiguity in SfM from a noisy flow field computed from two frames can be completely eliminated by using noise free inertial rate data. Secondly, we show that the admissible solution space for SfM from noisy feature correspondences can be reduced by using inertial data.

1. INTRODUCTION

Ambiguities in 3D motion recovery from noisy flow fields have been reported by many researchers [1, 2, 3]. Because of the observation noise, under many circumstances, multiple admissible camera motion and scene structure interpretation exist for a given noisy flow field. One dominant ambiguity arises from the similarity between the flow fields generated by translation parallel to the image plane and associated rotation [3] when the size of the field of view is small. Since this translation-rotation confusion is inherent, actively fixating on the focus of expansion is suggested in [3] to keep the lateral translation as small as possible such that this ambiguity can be attenuated. However, active control of camera motion is not available in many applications, such as passive navigation. One alternative way to eliminate this translation-rotation confusion is to exploit inertial rate data obtained from camera-fixed gyroscope. By using the inertial rate data, the flow field generated by pure rotation can be roughly predicted and hence the flow field generated by pure camera translation can be computed by subtracting the rotational flow field from the original observed flow field. Based on this noisy pure translational flow field, camera translation and 3D scene

Prepared through collaborative participation in the Advanced Sensors Consortium (ASC) sponsored by the U.S. Army Research Laboratory under the Federated Laboratory Program, Cooperative Agreement DAAL01-96-2-0001.

structure can be robustly estimated [4]. The translation rotation confusion can be completely eliminated when the inertial data are accurate.

Although the inherent ambiguities in SfM from a noisy flow field have been analyzed by many researchers, whether the ambiguities can be removed by using multiple frames has not been widely addressed, i.e. the uniqueness of solution for SfM using noisy feature correspondences has not been established yet. Although many batch and recursive estimators have been proposed for SfM using a set of noisy feature correspondences from long sequences, there is no guarantee that these methods can find a robust solution. In this paper, we also show that multiple admissible solutions exist for some sequences when the size of the field of view is small under certain measurement noise level and inertial rate data can be used to reduce the size of the admissible solution space.

2. TRANSLATION-ROTATION AMBIGUITY ELIMINATION FOR SFM FROM NOISY FLOW FIELD

The flow field is related to the 3D camera motion and feature structures by the well known flow equation [5].

$$\begin{cases} u(x, y) = u_t(x, y) + u_r(x, y) \\ v(x, y) = v_t(x, y) + v_r(x, y) \end{cases} \quad (1)$$

where (u_t, v_t) and (u_r, v_r) are, respectively, the translational and rotational components of the flow field and they are given by

$$\begin{cases} u_t(x, y) = (xt_z - t_x) \frac{1}{z(x, y)} \\ v_t(x, y) = (yt_z - t_y) \frac{1}{z(x, y)} \end{cases} \quad (2)$$

and

$$\begin{cases} u_r(x, y) = xy\omega_x + (1 + x^2)\omega_y + y\omega_z \\ v_r(x, y) = (1 + y^2)\omega_x - xy\omega_y - x\omega_z \end{cases} \quad (3)$$

where (t_x, t_y, t_z) is camera translation and $(\omega_x, \omega_y, \omega_z)$ is camera rotation. In practical applications, the observed flow field is corrupted by additive noise.

$$\begin{cases} \tilde{u}(x, y) = u(x, y) + n_u \\ \tilde{v}(x, y) = v(x, y) + n_v \end{cases} \quad (4)$$

where n_u and n_v are white noise with covariance σ_u^2 and σ_v^2 , respectively. Multiple admissible camera motion and scene structure interpretation exist for a given noisy flow field under many circumstances. The dominant ambiguity, translation rotation confusion, arises from the similarity between parallel translational flow fields associated rotational flow fields [3] when the size of the field of view is small. The availability of inertial rate data makes it possible to eliminate this translation rotation confusion, hence greatly reducing the ambiguities in 3D motion recovery from noisy flow fields. In practice, the inertial rate data are also corrupted by noise. The measurement equations are

$$\begin{cases} \tilde{\omega}_x = \omega_x + n_x \\ \tilde{\omega}_y = \omega_y + n_y \\ \tilde{\omega}_z = \omega_z + n_z \end{cases} \quad (5)$$

where n_x , n_y and n_z are also additive white noise with covariance σ_x^2 , σ_y^2 and σ_z^2 , respectively. By using the inertial data, the flow field generated by camera rotation can be predicted as

$$\begin{aligned} \tilde{u}_r(x, y) &= xy\tilde{\omega}_x - (1+x^2)\tilde{\omega}_y + y\tilde{\omega}_z \\ &= u_r(x, y) + n_{u,r} \end{aligned} \quad (6)$$

and similarly we have

$$\tilde{v}_r(x, y) = v_r(x, y) + n_{v,r} \quad (7)$$

where

$$\begin{cases} n_{u,r} = xy n_x - (1+x^2)n_y + yn_z \\ n_{v,r} = (1+x^2)n_x - xy n_y + xn_z \end{cases}$$

are the prediction errors of the rotational flow field with covariances given by

$$\begin{cases} \sigma_{u,r}^2 = (xy)^2\sigma_x^2 + (1+x^2)^2\sigma_y^2 + y^2\sigma_z^2 \\ \sigma_{v,r}^2 = (xy)^2\sigma_y^2 + (1+y^2)^2\sigma_x^2 + x^2\sigma_z^2 \end{cases} \quad (8)$$

A noisy pure translational flow field can be computed by subtracting the predicted rotational flow field from the originally observed flow field.

$$\begin{aligned} \tilde{u}_t(x, y) &= \tilde{u}(x, y) - \tilde{u}_r(x, y) \\ &= u_t(x, y) + n_{u,t} \end{aligned} \quad (9)$$

and similarly, we can have

$$\tilde{v}_t(x, y) = v_t(x, y) + n_{v,t} \quad (10)$$

where $n_{u,t}$ and $n_{v,t}$ are the observation noise for the pure translational flow field with covariances $\sigma_{u,t}^2$ and $\sigma_{v,t}^2$, respectively and

$$\begin{cases} \sigma_{u,t}^2 = \sigma_u^2 + (xy)^2\sigma_x^2 + (1+x^2)^2\sigma_y^2 + y^2\sigma_z^2 \\ \sigma_{v,t}^2 = \sigma_v^2 + (xy)^2\sigma_y^2 + (1+y^2)^2\sigma_x^2 + x^2\sigma_z^2 \end{cases} \quad (11)$$

Once the noisy translational flow field and associated measurement noise characterization is obtained, translation and 3D scene structure can be robustly estimated [4] when the inertial data are accurate, i.e. the inertial data observation noise covariances (σ_x^2 , σ_y^2 , σ_z^2) are small. Hence, the effects of translation rotation confusion are considerably reduced.

3. AMBIGUITY REDUCTION FOR SFM FROM MULTIPLE VIEWS

In this section, we synthesize an image sequence and show that at least two admissible solutions exist for this image sequence. The admissibility of a solution \mathbf{x} is measured by the likelihood function of the observation \mathbf{h}_0 given \mathbf{x} .

$$\alpha(\mathbf{x}, \mathbf{h}_0) = \ln(\mathbf{h}_0|\mathbf{x}) \quad (12)$$

When the measurement noise is Gaussian, $\alpha(\mathbf{x}, \mathbf{h}_0)$ is proportional to the weighted-square-cost $C(\mathbf{x}, \mathbf{h}_0)$. $C(\mathbf{x}, \mathbf{h}_0)$ is given by

$$C(\mathbf{x}, \mathbf{h}_0) = [\mathbf{h}_0 - \mathbf{h}(\mathbf{x})]^T \mathbf{R}^{-1} [\mathbf{h}_0 - \mathbf{h}(\mathbf{x})] \quad (13)$$

where $\mathbf{h}(\mathbf{x})$ is the measurement equation and \mathbf{R} is the covariance matrix of the measurement noise.

3.1. Existence of Multiple Solutions

Two 3-D coordinate systems are involved to describe the camera motion and feature point structures. One is an inertial world coordinate system and the other is a camera coordinate system which has its origin overlapping with the focal point of the camera. These two coordinate systems are coincident when the first image frame in a sequence is captured. For convenience, we call the world system I and the camera system C . When camera moves, I remains fixed on the ground while C moves with the camera. Hence, a static scene point will have a varying coordinate in C when the camera moves. The perspective projection of the feature point on the image plane forms the image of that point. This camera model we used in our approach is very similar to the one used in [6, 7].

In our simulation, we employed a virtual camera with focal length 2560 pixels to produce the video sequences with image size 512×512 . The camera translates along X axis with constant velocity -0.2 per second and rotates about Y axis with constant angular velocity 0.1 radian per second. Fourteen feature points are tracked through 57 image frames. Additive white Gaussian noises with zero mean, 2 pixels standard deviation are added to the synthesized image sequences. The structure parameters are set as shown in Table 1. For each feature point, (u, v) is the position of the perspective projection of the feature point onto the image plane in the first image frame and α is the true depth of the feature point. It is easy to check that these feature points are non-coplanar and they are not on a cone surface containing the center of projection. Hence, they satisfy the assumption of the uniqueness theorem in [8]. In [6], the task of estimating camera motion and scene structure is cast into an extended Kalman filtering (EKF) framework. Camera motion and scene structure are set to be the state parameters. They can be estimated given the 2D projections of the feature points in the image sequence in a recursive way. By using the EKF estimator

Table 1. Ground truth of 3D structure

Features	u	v	α	α_t	α_f
1	-0.10	0.05	0.71	0.7206	2.4283
2	-0.097	0.12	0.45	0.4553	4.7040
3	-0.1	0.135	0.7	0.7219	2.4248
4	0.1	-0.15	0.5	0.5113	3.9169
5	0.2	0.1	0.6	0.6105	2.8635
6	0.1	0.1	0.9	0.8908	1.3974
7	0.14	-0.1	0.3	0.3035	6.4214
8	0.07	0.12	0.85	0.8561	1.5853
9	0.071	-0.135	0.34	0.3496	5.9230
10	-0.1	0.15	0.4	0.3964	5.4324
11	0.1	0.25	0.5	0.4910	3.9650
12	-0.2	0.1	0.8	0.7968	1.9762
13	-0.14	0.1	0.3	0.3029	7.0546
14	0.043	0.129	1.0	1.0000	1.0000

with good initial state values and appropriate dynamic covariance matrix structure, one solution near ground truth could be obtained. The estimates of depth of feature points are $\{\alpha_t\}$ which are shown also in Table 1. We denote this true solution as S_t . We can also obtain another solution for the same observation. The associated estimated depth values are $\{\alpha_f\}$, listed in the last column in Table 1. We denote this false solution as S_f . We could see that this set of solution is far away from the ground truth. Let us check the admissibility of these two solutions. Since we assume the measurement noise is Gaussian, the likelihood function is proportional to the weighted-square-cost. The resulting weighted-square-cost of S_t and S_f are 2553.9 and 2255.1, respectively. We could see that it is hard to distinguish between these two solutions. The false solution even gives a lower cost than the true one. Hence, this example illustrates that multiple admissible solutions exist for SfM using noisy feature correspondence over image sequences. Here, admissible solutions are referred to the solutions which can produce comparable cost with the true solution. This example also gives an insight on the reverse-depth phenomenon in [6] where depth reversal of the structure and motion were occasionally experienced when the size of field of view is small. The existence of another solution plays a key role when the recursive estimator proposed in [6] converges to reversed depth.

3.2. Ambiguity function analysis

To visualize the existence of multiple admissible solution in the solution space, ambiguity function analysis is performed for the above scene structure and camera motion. The ambiguity function is often used in error analysis [9] for maximum likelihood estimation or least-square estimation. Assume that the measurement model is given by

$$\mathbf{z} = \mathbf{h}(\mathbf{x}, \mathbf{v}) \quad (14)$$

where \mathbf{z} is the observation and \mathbf{x} is the state vector and \mathbf{v} is the noise. Then if $\hat{\mathbf{x}}$ is a maximum likelihood estimate, the

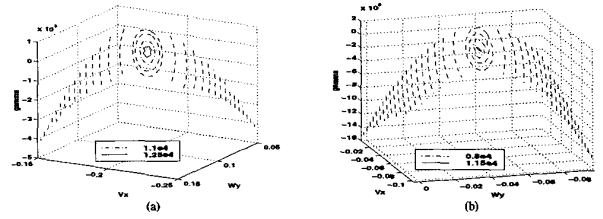


Fig. 1. Ambiguity functions around two admissible solutions. (a) is the ambiguity function around S_t and (b) is the ambiguity function around S_f .

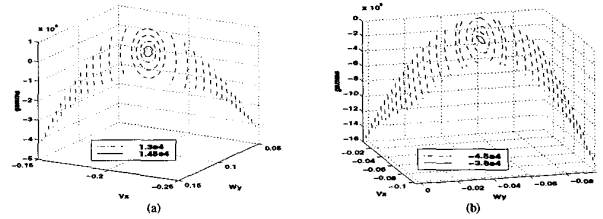


Fig. 2. Ambiguity functions around two estimates with inertial data. (a) is the ambiguity function around S_t and (b) is the ambiguity function around S_f .

ambiguity function around the true state \mathbf{x}_t , $\gamma(\mathbf{x}_t, \hat{\mathbf{x}})$, is defined as the average of the log likelihood function by

$$\gamma(\mathbf{x}_t, \hat{\mathbf{x}}) = \int \ln[p(\mathbf{z}|\hat{\mathbf{x}})]p(\mathbf{z}|\mathbf{x}_t)dz \quad (15)$$

where $p(\mathbf{z}|\mathbf{x})$ is the likelihood function of \mathbf{z} given \mathbf{x} . When \mathbf{v} is normal with zero mean and covariance \mathbf{R} ,

$$\gamma(\mathbf{x}_t, \hat{\mathbf{x}}) = \frac{1}{2}(-\ln[(2\pi)^N |\mathbf{R}|] - N - [\mathbf{h}(\mathbf{x}_t) - \mathbf{h}(\hat{\mathbf{x}})]^T \mathbf{R}^{-1} [\mathbf{h}(\mathbf{x}_t) - \mathbf{h}(\hat{\mathbf{x}})]) \quad (16)$$

The shape of ambiguity function can give a direct feel of the solution space of a particular estimation problem. We compute the ambiguity functions around S_t and S_f and the contour lines of the associated ambiguity functions are shown in Figure 1. It can be seen that there are two peaks with comparable maximum values around the two solutions. One way to remove the false solution S_f is to use the inertial rate data described in Section 2 and use it as a direct measurement of the inter-frame rotation in the extended Kalman recursive estimator [6]. We can obtain a solution similar to S_t and the resulting cost function is 3096. The cost function of S_t increases dramatically to 51763. Hence, S_f is not admissible any longer once the inertial data are available. We also compute the ambiguity functions around S_t and S_f in this case and the contour lines of the two ambiguity functions are shown in Figure 2. The maximum value of the ambiguity function around S_f is negative and it will be not accepted as a solution. Hence the use of inertial rate data can reduce the number of admissible solutions.

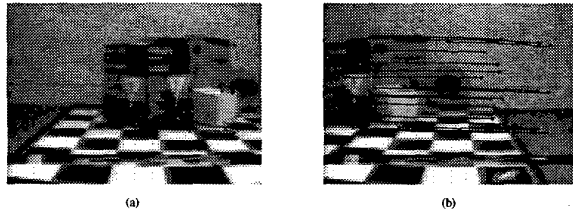


Fig. 3. Feature points and their trajectories tracked through a translational sequence. The bar-code containing inertial rate data can be seen at the bottom of both frames.

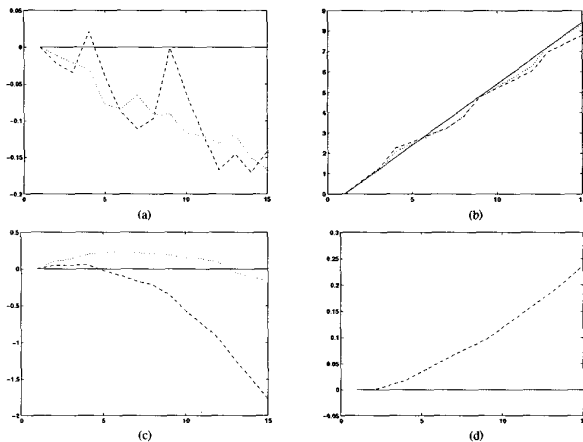


Fig. 4. Motion estimates from a real image sequence. (a), (b) and (c) show camera translation (inch). (e) shows the camera rotation about X axis (radian).

3.3. Experiments using real image sequences

A video sequence containing inertial data was used to test our algorithm. The video camera we used captures 25 image frames per second with a field of view (FOV) 46×34 (degrees). The size of the output image is 320×240 (pixels). Feature points are detected in the first frame and tracked automatically through the sequence.

When this sequence was captured, the camera was moved on a straight train track with approximate constant velocity 11.25 inches per second. Figure 3(a) is the first frame of this sequence with labeled feature points and Figure 3(b) is the last frame of the sequence with feature trajectories. In Figure 3, the bar-code containing inertial information can be seen at the bottom of the image frames. Figure 4 shows the motion estimates obtained using the method proposed in [6] (dashed lines) and our approach with inertial data (dotted lines). The ground-truth of the motion parameters is also shown using solid lines. In Figure 4, (a), (b) and (c) show camera translation in inch and (e) shows the camera rotation about X axis in radian. It can be seen that the motion estimates obtained using the inertial data are more accurate than the estimates from the method described in [6]. We also

computed the weighted-square-cost of both solutions. Surprisingly, if only image measurement is taken in account, the one further away from the ground truth has a lower cost of 827 than the other which has a cost of 911. This also shows the existence of multiple solutions. Of course, once inertial data were added into observation, the false solution had a much higher cost than that of the true solution. Hence, only the true solution is accepted as a valid solution.

4. CONCLUSIONS

Inertial rate data can be used to eliminate the translation rotation confusion in SfM from the noisy flow field. Multiple admissible solutions exist for SfM from feature correspondences tracked through long image sequences and the size of the admissible solution space can be reduced by using the inertial rate data.

5. ACKNOWLEDGMENT

The authors would like to thank Sanders, A Lockheed Martin Company for providing video data used in this paper.

6. REFERENCES

- [1] G. Adiv, "Inherent ambiguities in recovering 3-d motion and structure from a noisy flow field," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **11**, pp. 477-489, May 1989.
- [2] G. Young and R. Chellappa, "Statistical analysis of inherent ambiguities in recovering 3-d motion from a noisy flow field," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **14**, pp. 995-1013, October 1992.
- [3] K. Daniilidis and H. Nagel, "The coupling of rotation and translation in motion estimation of planar surfaces," in *IEEE Computer Vision and Pattern Recognition, New York, NY*, pp. 188-193, 1993.
- [4] D. Lawton, "Processing translational motion sequences," *Computer Vision, Graphics and Image Processing* **22**, pp. 116-144, April 1983.
- [5] G. Adiv, "Determining 3-d motion and structure from optical flow generated by several moving objects," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **7**, pp. 384-401, July 1985.
- [6] A. Azarbayejani and A. Pentland, "Recursive estimation of motion, structure, and focal length," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **17**, pp. 562-575, 1995.
- [7] R. Szeliski and S. Kang, "Recovering 3d shape and motion from image streams using non-linear least squares," in *IEEE Computer Vision and Pattern Recognition, New York, NY*, pp. 752-753, 1993.
- [8] R. Tsai and T. Huang, "Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **6**, pp. 13-27, January 1984.
- [9] F. Schweppe, *Uncertain Dynamic Systems*, Prentice-Hall, Englewood Cliffs, New Jersey, 1973.