

Recursive Structure and Motion Estimation from Noisy Uncalibrated Video Sequences

D. Q. Huynh A. Heyden

School of Computer Science and Software Engineering, The University of Western Australia, AUSTRALIA

School of Technology and Society, Malmö University, SWEDEN

du@csse.uwa.edu.au heyden@ts.mah.se

Abstract

This paper builds on a novel framework of hybrid matching constraints for estimation of structure and recovery of camera focal length and motion, combining the advantages of both discrete and continuous methods. Our recursive method can deal with both image noise and outliers. The system is an extension of the epipolar hybrid matching constraints in conjunction with a simple structure estimation scheme using standard triangulation. The extension enables the system to deal with varying focal length of the camera. The structure obtained from some previous image frames is used to improve estimates of the camera focal length and motion for the current image frame. These are, in turn, used to refine the structure. Finally, a RANSAC outlier rejection scheme is employed to reject outlier tracks, inevitably obtained from any tracker. The performance of the proposed system is demonstrated on simulated experiments.

1. Introduction

As with any parameter estimation problem that involves a sequence of observations, there are two broad categories of approaches: estimate the parameters in batch where observations from all image frames are used simultaneously; or estimate the parameters for the current state of the system based on the estimates from the previous state(s) and the new observation. Although more images need to be dealt with, the advantages of working with video sequences are twofold: 1) image feature point matching can be replaced by image feature point tracking; 2) camera motion and scene structure can be recovered recursively. The second advantage is particularly attractive as recursive approaches are often computationally more efficient and can easily be adapted to real-time systems. One of the ear-

lier research work in designing a recursive solution for 3D motion estimation from video sequences is that of Broida *et al.* [2]. They employ an iterated extended Kalman filter to estimate the motion and structure parameters that constitute the state vector of the system. Azarbajani and Pentland [1] extend their method to include the estimation of the camera focal length, while Soatto [13] imposes metric constraints on the state space so as to isolate the models for 3D structure and those for 3D motion. Gallego *et al.* [5] employ the Kalman filter to recursively update the 4×4 homography matrix that is required for upgrading the scene structure from projective to Euclidean.

Matching constraints provide important conditions on the geometry of the perspective projection to which the structure and motion parameters must conform. In [15], Triggs provides a detailed analysis of the bilinear, trilinear, and quadrilinear constraints. With image features undergoing small displacements between consecutive frames, matching constraints can be expressed in terms of image feature point positions and velocities (*e.g.*, see [10]). Recently, Nyberg and Heyden [12] and Heyden *et al.* [7] devise the hybrid matching constraints (HMC) for video sequences. These constraints can be viewed as analogous to the epipolar and trifocal constraints for the discrete case. With the assumption that the camera is calibrated, Heyden *et al.* [12] use these constraints to obtain the update for the current motion estimate of the camera linearly. Their HMC is then fused with a continuous-discrete extended Kalman filter for the state estimation of scene structure recursively.

In this paper, we adopt the HMC derived in [7, 12] and extend it to the case of uncalibrated camera and variable focal length. We incorporate the RANSAC paradigm to deal with outliers that arise in the feature tracking process and we analyze the HMC to determine the minimum number of correctly tracked image features required for recovering the structure, intrinsic and motion parameters of the camera.

2. Problem formulation

Under the perspective projection using the pinhole camera model, a given scene point $\tilde{\mathbf{X}} = (\mathbf{X}^\top, 1)^\top = (X, Y, Z, 1)^\top$ and its image point $\tilde{\mathbf{x}} = (\mathbf{x}^\top, 1)^\top = (x, y, 1)^\top$ satisfy the following relationship:

$$P\tilde{\mathbf{X}} = K [R \quad -\mathbf{b}] \tilde{\mathbf{X}} = \lambda\tilde{\mathbf{x}}. \quad (1)$$

The projection matrix P is $\mathbb{R}^{3 \times 4}$; the camera matrix K , if the camera's principal point is known, can be written as $\text{diag}(f, f, 1)$; λ is an unknown scalar; the rotation matrix $R \in SO(3)$ and translation vector $\mathbf{b} \in \mathbb{R}^3$ relate the 3D coordinate system of the camera with that in the scene. There are effectively two linear constraints that relate $\tilde{\mathbf{x}}$ and \mathbf{X} .

Given a static scene observed by a moving uncalibrated camera, the 3D structure of the scene is invariant but the intrinsic and extrinsic parameters of the camera are all continuous functions of time (or frame number), *i.e.*, our frame-dependent unknowns are $R(t)$, $\mathbf{b}(t)$, and $f(t)$, which, for simplicity, can be written as R_t , \mathbf{b}_t , and f_t , for $t = 0, \Delta t, 2\Delta t, \dots$. Let the estimated rotation, baseline, and focal length of the camera at frame t be R_t , \mathbf{b}_t , and f_t . Using the first two terms of the Taylor expansion, we can write these frame-dependent parameters at frame $t + \Delta t$ as

$$R_{t+\Delta t} \approx R_t + \Delta R_t \Delta t = (I + [\mathbf{w}_t]_\times \Delta t) R_t, \quad (2)$$

$$\mathbf{b}_{t+\Delta t} \approx \mathbf{b}_t + \mathbf{d}_t \Delta t, \quad (3)$$

$$K_{t+\Delta t} \approx K_t (I + \Delta K_t \Delta t), \quad (4)$$

where I denotes the identity matrix; \mathbf{w}_t is the vector that encapsulates the unknown angular velocity of R_t ; $[\cdot]_\times$ denotes the skew-symmetric matrix formed by the vector concerned; vector \mathbf{d}_t is the change of the baseline \mathbf{b}_t ; and $K_t = \text{diag}(f_t, f_t, 1)$ is the estimated camera matrix at frame t . To determine these frame-dependent parameters at frame $t + \Delta t$, it is then necessary to estimate the 7 frame-dependent parameter updates: $\Delta f_t \in \mathbb{R}$; $\mathbf{d}_t, \mathbf{w}_t \in \mathbb{R}^3$.

3. Our method

The *epipolar hybrid matching constraints* proposed in [7, 12] is analogous to the epipolar constraint but involves taking three images into account (an extra image is required for the image corner displacements or velocities). For the *trifocal hybrid matching constraints*, four images are used. In this paper, we will focus only on the linear epipolar hybrid matching constraints that involve looking at three images at a time.

3.1 Epipolar hybrid matching constraints

Without loss of generality, we fix the global 3D coordinate system at the optical centre of the camera at

frame 0. The perspective projections for images at frames 0, t , and $t + \Delta t$ are then

$$K_0 \mathbf{X} = \lambda_0 \tilde{\mathbf{x}}_0, \quad (5)$$

$$K_t [R_t \quad -\mathbf{b}_t] \tilde{\mathbf{X}} = \lambda_t \tilde{\mathbf{x}}_t, \quad (6)$$

$$K_{t+\Delta t} [R_{t+\Delta t} \quad -\mathbf{b}_{t+\Delta t}] \tilde{\mathbf{X}} = \lambda_{t+\Delta t} \tilde{\mathbf{x}}_{t+\Delta t}. \quad (7)$$

We adopt (2)-(4) and the Taylor approximations:

$$\lambda_{t+\Delta t} \approx \lambda_t + \Delta \lambda_t \Delta t, \quad \tilde{\mathbf{x}}_{t+\Delta t} \approx \tilde{\mathbf{x}}_t + \tilde{\mathbf{u}}_t \Delta t, \quad (8)$$

where $\tilde{\mathbf{u}}_t = (\cdot, \cdot, 0)^\top$ is the displacement of the image corner $\tilde{\mathbf{x}}_t$. By performing the following change of coordinate system:

$$\tilde{\mathbf{x}}'_0 \equiv K_0^{-1} \tilde{\mathbf{x}}_0, \quad \tilde{\mathbf{x}}'_t \equiv K_t^{-1} \tilde{\mathbf{x}}_t, \quad \tilde{\mathbf{u}}'_t \equiv K_t^{-1} \tilde{\mathbf{u}}_t. \quad (9)$$

we obtain

$$\begin{bmatrix} R_t \tilde{\mathbf{x}}'_0 & \tilde{\mathbf{x}}'_t & \mathbf{0} & \mathbf{b}_t \\ ([\mathbf{w}_t]_\times + \Delta K_t) R_t \tilde{\mathbf{x}}'_0 & \tilde{\mathbf{u}}'_t & \tilde{\mathbf{x}}'_t & \mathbf{d}_t + \Delta K_t \mathbf{b}_t \end{bmatrix} \begin{bmatrix} -\lambda_0 \\ \lambda_t \\ \Delta \lambda_t \\ 1 \end{bmatrix} = \mathbf{0}. \quad (10)$$

Note that (10) is not the same as the constraint derived in [7, 12] because of the presence of the ΔK terms in the matrix. However, we may proceed with the same analysis that the matrix in (10) has rank equal to 3 and that all of its 4×4 minors must vanish identically. Linear constraints for the 7 parameter updates can be obtained by appropriate selections of rows that compose these minors. It is straightforward to verify that the number of independent linear constraints is 2 and that if $N \geq 4$ correctly tracked corners are available, the 7 parameter updates can be estimated using least squares. This is different from the case in [7, 12], since only 3 corner tracks were required there.

3.2 Refinement by minimizing reprojection errors

The refinement step is designed to improve the estimates $f_{t+\Delta t}$, $R_{t+\Delta t}$, and $\mathbf{b}_{t+\Delta t}$ obtained from the parameter updates using HMC. Firstly, we obtain the scene structure \mathbf{X} via triangulation and advance the frame number by Δt , *i.e.*, we set $t \leftarrow t + \Delta t$. Next, we define

$$R_t^+ = \exp([\delta \mathbf{w}_t]_\times) R_t; \quad \mathbf{b}_t^+ = \mathbf{b}_t + \delta \mathbf{d}_t; \quad K_t^+ = K_t (I + \delta K_t), \quad (11)$$

where $\delta \mathbf{w}_t, \delta \mathbf{d}_t \in \mathbb{R}^3$ and $\delta K_t = \text{diag}(\delta f, \delta f, 0)$ are the refinement updates that need to be estimated. Those symbols having a '+' superscript denote that they are refined estimates to be computed over those obtained from the HMC. Substituting (11) into $\lambda_t^+ \tilde{\mathbf{x}}_t = K_t^+ [R_t^+ \quad -\mathbf{b}_t^+] \tilde{\mathbf{X}}$ and using $R_t^+ \approx (I + [\delta \mathbf{w}_t]_\times) R_t$ yields

$$\begin{aligned} \lambda_t^+ \tilde{\mathbf{x}}'_t &\approx R_t \mathbf{X} - \mathbf{b}_t + \delta K_t R_t \mathbf{X} - \delta K_t \mathbf{b}_t + [\delta \mathbf{w}_t]_\times R_t \mathbf{X} - \delta \mathbf{d}_t \\ \Rightarrow R_t \mathbf{X} - \mathbf{b}_t - \lambda_t^+ \tilde{\mathbf{x}}'_t &:= e \approx [R_t \mathbf{X}]_\times \delta \mathbf{w}_t + \delta \mathbf{d}_t - \delta K_t R_t \mathbf{X} \\ &\quad + \delta K_t \mathbf{b}_t, \quad (12) \end{aligned}$$

where $\tilde{\mathbf{x}}'$ is defined in (9). Note that e can be interpreted as the reprojection error. We call (12) the refinement based on reprojection constraints. The refinement updates, $\delta\mathbf{w}_t$, $\delta\mathbf{d}_t$, and δf , can be estimated using least squares if $N \geq 4$ correctly tracked corners are available. The scene structure \mathbf{X} can be further refined using the new estimates of \mathbf{w}_t , \mathbf{d}_t , and Δf_t .

3.3 The recursive procedure

Initialization: Given two images at frames 0 and t , (i) compute the fundamental matrix F using the 8 point algorithm; (ii) construct K_0 and K_t by applying a camera self-calibration algorithm, *e.g.*, [11], that estimates f_0 and f_t ; (iii) compute the essential matrix $E = K_t F K_0$; (iv) set $R_0 = I$, $\mathbf{b}_0 = \mathbf{0}$, and compute R_t and \mathbf{b}_t from E via a method described in, *e.g.*, [6]; (v) compute an initial estimate, $\mathbf{X}^{(i)}$, of each of the identified inliers; (vi) compute $\mathbf{x}_0^{(i)} = K_0^{-1} \mathbf{x}^{(i)}$ and $\mathbf{x}_t^{(i)} = K_t^{-1} \mathbf{x}^{(i)}$; (vii) create an inlier table to maintain all the inliers identified by RANSAC at each image frame.

We scale \mathbf{b}_t to unit magnitude and the baselines for subsequent video frames are defined relative to this scale. In the presence of outliers, step (i) should be wrapped inside a RANSAC loop.

Recursive procedure:

- (i) For each image corner, $\tilde{\mathbf{x}}_{t+\Delta t}^{(i)}$, apply the following change of coordinate system: $\tilde{\mathbf{x}}_{t+\Delta t}^{(i)} = K_t^{-1} \tilde{\mathbf{x}}_{t+\Delta t}^{(i)}$. Note that K_t is involved in the matrix multiplication here rather than $K_{t+\Delta t}$ which has yet to be computed.
- (ii) Sample 4 image corners from (i) and estimate Δf_t , \mathbf{w}_t , and \mathbf{d}_t (Section 3.1).
- (iii) Construct $K_{t+\Delta t}$, $R_{t+\Delta t}$, and $\mathbf{b}_{t+\Delta t}$ via (2)-(4). Use these matrices and the estimated \mathbf{X} to compute the reprojection errors of all image corners. These errors are used by RANSAC as a measure of ‘well-being’ of the sample chosen in Step (ii).
- (iv) After all the inliers have been identified by RANSAC (this requires some pre-defined threshold values and knowledge about the percentage of outliers in the data), update the inlier table and recompute $K_{t+\Delta t}$, $R_{t+\Delta t}$, and $\mathbf{b}_{t+\Delta t}$ (Section 3.1) using *all* the inliers. Note that as the HMC involves looking at 3 images (at frames 0, t , and $t + \Delta t$) simultaneously, only image corners that are inliers in all these three images should be used in the computation.
- (v) Compute the refinement updates (Section 3.2) using all the inliers for frames 0, t , and $t + \Delta t$; triangulate for \mathbf{X} . Note that t becomes $t + \Delta t$ after this step.
- (vi) Recompute $\tilde{\mathbf{x}}_t^{(i)}$, for all i , using the newly estimated K_t , *i.e.*, set $\tilde{\mathbf{x}}_t^{(i)} = K_t^{-1} \tilde{\mathbf{x}}_t^{(i)}$.
- (vii) Loop back to Step (i) of the recursive procedure for

the next image frame.

4. Experimental results and discussion

Two of the many synthetic experiments conducted are reported here. In the first experiment the camera underwent a simpler and smoother trajectory (*i.e.*, with small amount of camera rotations and change of depths) whereas in the second experiment more camera rotations and changes of depth were involved. In both experiments, small Gaussian noise was added to the coordinates of all image corners to simulate inlier noise; a small number of them were perturbed by a displacement that was about the size of a 5×5 tracking window, over a small number of frames, to simulate outliers.

We carried out the method described in the previous section and evaluated the errors of our estimated parameters. Figs. 1 and 2 show the error plots of these parameters over the video frame number for the two experiments. Symbols with an overhead bar denote the true values and symbols with a hat denote the estimated values of the parameters. In the figure, $\epsilon_{\mathbf{X}}$ denotes the average Euclidean distance between the true and estimated scene points after they have been aligned by an estimated similarity transformation. Only the inliers were included in the error computation. The variable ϵ_R , $\epsilon_{\mathbf{b}}$, and ϵ_f are defined as follows: $\epsilon_R = \|\log(\hat{R}R^T)\|$; $\epsilon_{\mathbf{b}} = \|\bar{\mathbf{b}} - \hat{\mathbf{b}}\|/\|\bar{\mathbf{b}}\|$; $\epsilon_f = |(\bar{f} - \hat{f})/\bar{f}|$.

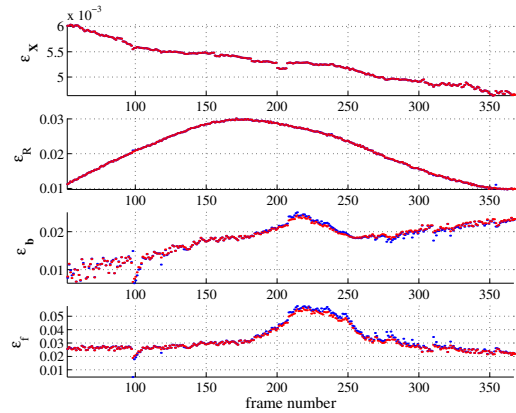


Figure 1. Error plot for experiment 1.

In both experiments, improvements to the estimated 3D structure are evident. Although the error seemed to increase slightly after frame 300 in experiment 2, the error at the last frame was still sufficiently small to be neglected. The errors of the estimated rotations for both experiments were very small. The maximum rotation error in experiment 1 was only 0.03 radians, *i.e.*, 1.72 deg. A possible explanation for the pattern of rota-

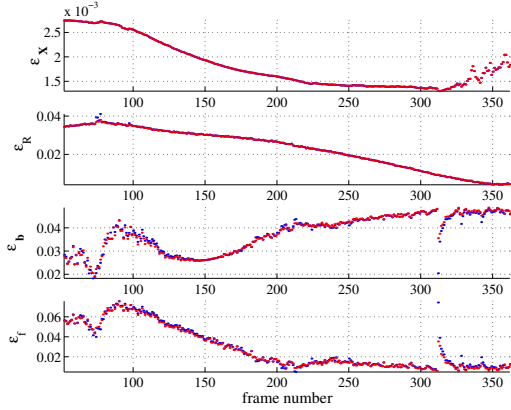


Figure 2. Error plot for experiment 2.

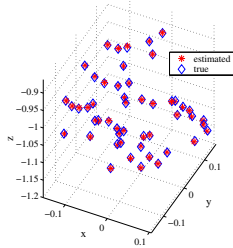


Figure 3. The estimated and true 3D structures at the final frame of experiment 2.

tion error in experiment 1 is the specific camera 3D trajectory in the simulation together with the combination of focal length and baseline errors. Experiment 1 had 8% outlier tracks, with average outlier track length of 10 video frames. Experiment 2 had 40% outlier tracks whose average length is 20 frames. In both experiments, all outlying tracks were successfully detected. The estimated 3D reconstruction for experiment 2 shows a perfect alignment with the true 3D reconstruction (Fig. 3). The reconstruction result for experiment 1 is similar.

One may notice that the error patterns for the baseline and focal length appear to be similar; This is not unexpected as an overestimate (resp. underestimate) of the focal length would require the baseline to be lengthened (resp. shortened) in order to minimize the average reprojection error, which is a constraint imposed in our method.

We have also conducted some preliminary experiments on real video data. As the current version of our algorithm can only deal with continuous tracks, it is a challenging task to capture long image sequences in which a sufficient number of good and continuous

tracks are present for testing the algorithm. One of the video sequences that we tested is the well known `hotel` image sequence (each image is 480 rows \times 512 columns in size) available on the CMU website [3]. We selected a very short portion of the image sequence (comprising frames 0, 40, 41, \dots , 50) and applied the SIFT keypoint detector [8] to all the images. A large skip from frame 0 to frame 40 was necessary for the initialization of the algorithm (see previous section). To construct the corner tracks, we modified the image matching program (written in C) provided on Lowe’s website to work on image sequences, i.e., SIFT keypoints from image frame 0 were matched with those in frames 40, 41, and so on. We also combined some corner tracks from the KLT tracker [9, 14] to obtain more continuous tracks (see Figure 4). It is obvious in the figure that quite a few outliers were present. The algorithm presented in [11] for estimating the focal lengths of two cameras requires the knowledge of the principal point for each image. Unfortunately, in the absence of ground truth information, we could only assume that the principal point was approximately at the centre of the image buffer¹. We note that if there is a large deviation between the true and the assumed principal points then it can result in large disparity errors and thus poor 3D reconstruction. From these noisy input corner tracks, the fundamental matrix was estimated from a RANSAC loop and the focal lengths were estimated to be 506.91 and 540.25 pixels, respectively for frames 0 and 40. Most of the outlying tracks were pruned away in the RANSAC loop.

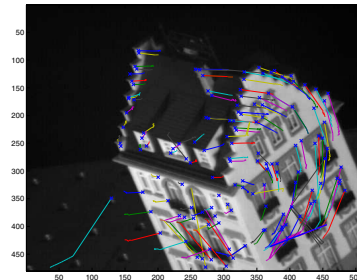


Figure 4. The CMU hotel image sequence with corner tracks superimposed.

The initial and final reconstructions were very similar and were not perfect. Skewness can be easily detected in the reconstruction (see Figure 5). We conclude

¹Although vanishing points can be used to estimate the principal point in an image, they must be at finite distance, preferably within the image frame boundary. For the `hotel` image sequence, the vanishing points are almost at infinity.

that, to fully test the method on real data, it is necessary that more continuous tracks that cover the entire image (note that the hotel model occupies mainly the lower right corner of each image) are available and that the system is well initialized. If only two images are used for the initialization stage, then the principal point for each image must be known. If three or more images are available, then other self-calibration techniques can also be employed (e.g. [4]).

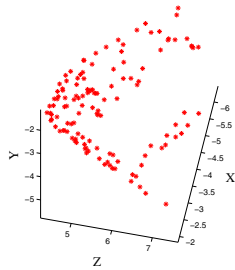


Figure 5. Reconstruction of the CMU hotel image sequence. It can be seen that the walls are not perfectly orthogonal to each other.

It is important that a reasonably good initial estimate of the focal length is obtained for frames 0 and t , since the estimated K_t matrix is then used for the change of image coordinate system. From our experiments, it seems that more complex camera trajectories can sometimes yield better parameter estimate than do simple camera trajectories, even in the presence of a higher proportion of outlier tracks. A possible explanation is that locally degenerate configurations of the camera are unlikely to occur for more complex trajectories. For long video sequences, one may be able to reduce numerical errors by continuously moving the global coordinate system closer to the current video frame.

5. Conclusion and future work

We have presented a recursive method for structure and motion recovery from uncalibrated video sequences. The method is an extension of [7, 12], which works only for the calibrated case with no outliers. So far our experiments on synthetic data show that the method is effective in that the errors on the estimated focal length, baseline, rotation, and the recovered 3D structure are all very low. By incorporating the RANSAC paradigm, we were able to successfully identify and eliminate all the outlier tracks from the parameter computation. Our method has not yet been designed to handle missing data, e.g., image corner tracks disappear or new image corner tracks emerge part-way

through the video sequences. However, it would be straightforward to incorporate this into a later version of our method, which we also intend to test on more real video sequences.

References

- [1] A. Azarbayejani and A. P. Pentland. Recursive Estimation of Motion, Structure, and Focal Length. *IEEE Trans. on PAMI*, 17(6):562–575, 1995.
- [2] T. J. Broida, S. Chandrashekar, and R. Chellappa. Recursive 3-D Motion Estimation from a Monocular Image Sequence. *IEEE Trans. on Aerospace and Electronic Systems*, 26(4):639–656, 1990.
- [3] CMU Image Database website. <http://vasc.ri.cmu.edu/idb/html/motion/index.html>.
- [4] O. D. Faugeras, Q. T. Luong, and S. J. Maybank. Camera Self-Calibration: Theory and Experiments. In G. Sandini, editor, *Proc. ECCV*, pages 321–334, May 1992.
- [5] G. Gallego, J. I. Ronda, A. Valdés, and N. García. Recursive Camera Autocalibration with the Kalman Filter. In *Proc. ICIP*, 2007.
- [6] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2004.
- [7] A. Heyden, F. Nyberg, and O. Dahl. Recursive Structure and Motion Estimation based on Hybrid Matching Constraints. In *SCIA*, 2007.
- [8] D. G. Lowe. Object Recognition from Local Scale-Invariant Features. In *Proc. ICCV*, pages 1150–1157, Corfu, Greece, Sep 1999.
- [9] B. D. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *IJCAI*, pages 674–679, 1981.
- [10] Y. Ma, J. Košecká, and S. Sastry. Linear Differential Algorithm for Motion Recovery: A Geometric Approach. *IJCV*, 36(1):71–89, Jan 2000.
- [11] G. N. Newsam, D. Q. Huynh, M. J. Brooks, and H.-P. Pan. Recovering Unknown Focal Lengths in Self-Calibration: An Essentially Linear Algorithm and Degenerate Configurations. In *ISPRS*, volume XXXI, part B3, commission III, pages 575–580, Jul 1996.
- [12] F. Nyberg and A. Heyden. Recursive Structure from Motion using Hybrid Matching Constraints with Error Feedback. In *Workshop on Dynamic Vision (held at ECCV'06)*, 2006.
- [13] S. Soatto. 3-D Structure from Visual Motion: Modeling, Representation and Observability. *Automatica*, 33(7):1287–1312, 1997.
- [14] C. Tomasi and T. Kanade. Detection and Tracking of Point Features. Technical Report CMU-CS-91-132, Carnegie Mellon University, Pittsburgh, PA, USA, Apr 1991.
- [15] B. Triggs. Matching Constraints and the Joint Image. In *Proc. ICCV*, pages 338–343, 1995.