

A Semi-direct Approach to Structure From Motion

Paolo Favaro
Washington University
Electrical Engineering
Campus Box 1127
St. Louis, MO 63130
fava@ee.wustl.edu

Hailin Jin
Washington University
Electrical Engineering
Campus Box 1127
St. Louis, MO 63130
hljin@ee.wustl.edu

Stefano Soatto
University of California, Los Angeles
Computer Science
Los Angeles, CA 90095 and
Washington University, St. Louis
soatto@cs.ucla.edu, soatto@ee.wustl.edu*

Abstract

Reconstructing three-dimensional structure and motion is often decomposed into two steps: point feature correspondence and three-dimensional reconstruction. This separation often causes gross errors since correspondence relies on the brightness constancy constraint that is local in space and time. Therefore, we advocate the necessity to integrate visual information not only in time (i.e. across different views), but also in space, by matching regions - rather than points - using explicit photometric deformation models. We present an algorithm that integrates 2D region tracking and 3D motion estimation into a closed loop based on an explicit geometric and photometric model, while detecting and rejecting outlier regions that do not fit the model. Our algorithm is recursive and suitable for real-time implementation. Our experiments show that it far exceeds the accuracy and robustness of point feature-based SFM algorithms.

1 Introduction

Structure from motion consists in recovering a three-dimensional model of the shape and motion of a scene from a number of (monocular) images obtained during motion. This problem is typically broken down into two: first establish point-to-point correspondence among different images using assumptions on the photometry of the scene (e.g. Lambertian reflection), then use image correspondence to infer three-dimensional geometry¹. In particular, we are interested in on-line estimation of shape and motion for the purpose of control (e.g. driving a vehicle in an unknown environment). In this scenario, we assume that images are

taken at adjacent instants in time while the viewer and/or the scene move smoothly, and impose the causal constraint that only images up to the current time t can be used to infer the estimate of shape and motion at time t .

When vision is to be used as a sensor in a closed control loop, delays in the estimates can have catastrophic consequences, which render the most popular schemes based on processing batches of views infeasible. This is not just a matter of computational speed: the process of collecting a number of views (say 3 or more), applying interest operators, establishing correspondence using statistical consensus algorithms, computing epipolar geometry and finally refining and densifying the estimates using multi-frame bundle adjustment would result in delays that cannot be reduced below a certain threshold that is too high for any practical control purpose such as driving or manipulation.

It can be argued that the problem of visual reconstruction is ill-posed unless a specific purpose for the estimates is specified, for instance in the form of a control task². In our view, the causal aspect of structure from motion (SFM), while fundamental, has received comparatively little attention in the Computer Vision community. Therefore, despite the wealth of efforts and results in the field of SFM, we believe that there remains the need for robust algorithms for estimating structure and motion under the causal constraint.

*This research is supported in part by ARO grant DAAD19-99-1-0139 and Intel grant 8029.

¹There are of course exceptions to this general scheme, as we discuss in section 1.2.

²Images depend upon the unknown geometry, photometry and dynamics of the scene as well as on the distribution of the light source; images alone are not sufficient to uniquely estimate all unknowns, thus resulting in an intrinsically ill-posed problem. While priors can be imposed (e.g. on photometry) to render the problem well-posed, they can never be validated from image data nor detected with statistical techniques: if one assumes that a specular object is Lambertian, her reconstruction algorithms will return a consistent estimate of the wrong shape. The situation changes if the scope of the inference is to specify a control task. For instance, in specifying a visual servo command the control signal can be specified directly from measurements on the image plane, even if the resulting model of the scene is not correct in the Euclidean sense, the task can be performed correctly (i.e. with asymptotically stable error).

1.1 Integration in space and time

In past research, we have found that the most delicate step in reconstructing structure and motion in real time is establishing point correspondences. Point matching relies on the photometric models that are local in space and time, and is therefore prone to mismatches, outliers, drift and other inconveniences that significantly impact the robustness of the overall system. Therefore, we advocate the necessity to integrate visual information not only in time (i.e. across different views), but also in space, by matching regions - rather than points - using explicit photometric deformation models.

Unfortunately, the deformation undergone by image irradiance functions as a consequence of rigid motion cannot be captured by a finite-dimensional model (see section 2). Therefore, we will not seek to model global deformations. Instead, we will choose a finite-dimensional parameterization of photometric deformations, and segment the image into regions that satisfy the model (as verified in a statistical hypothesis test). Visual information will then be integrated locally in space (within a region), and globally in time (within a rigid object), while occluding boundaries and specular reflections are detected explicitly as violating the hypothesis. Of course the size of the region will depend upon the maximum discrepancy from the model that we are willing to tolerate, and in general there will be a tradeoff between robustness (calling for larger regions) and accuracy (calling for smaller ones). Such a tradeoff can be addressed using information-theoretic tools [13]. In practice it is not necessary to cover the whole image with regions, since regions with small irradiance gradient do not impose shape constraints, and therefore significant speedups can be achieved.

One can view our effort as a step towards a dense representation of shape, moving from points to surfaces, with an explicit model of illumination. Indeed, we seek to integrate into a unified scheme photometry (feature tracking), dynamics (motion estimation) and geometry (point-wise reconstruction and surface interpolation). In particular, in our experimental assessment, we represent a piecewise smooth surface with a collection of rigidly connected planes supporting a radiance function that undergoes projective deformations. Spatial grouping allows a significant reduction of complexity, since points need not be detected and tracked individually.

1.2 Relation to previous work

The present work falls within the category of structure from motion (SFM), a field that encompasses a vast variety of research efforts, such as [2, 4, 5, 11, 14, 15, 16, 21, 22]. Of all the work in SFM, we consider in particular causal

estimation algorithms. A batch approach would obviously perform better, but at the expense of making the estimates useless when it comes to performing control actions such as manipulation, navigation or, more in general, real-time interaction where delays cannot be tolerated [12].

Since we integrate tracking and motion estimation, our work also relates to the large literature on image (2D) motion. However, most tracking schemes rely on point features and do not exploit feedback from higher levels. If the scene is a rigid collection of features that undergo the same rigid motion, this global constraint can be enforced by a feature tracker for robustness and precision. A small body of literature on so-called direct methods addresses this issue, for example [7, 19]. The basic idea is to use the same brightness constancy constraint equation that is used to estimate optical flow or feature displacement as an implicit measurement of an extended Kalman filter (EKF) that estimates motion parameters. Image motion is then integrated globally, as long as the brightness constraint is satisfied. The exact constraint, however, depends upon the shape of the scene, which is unknown. Most work in direct SFM ends up representing shape as a collection of points whose projections are subject to brightness constancy and undergo the same rigid motion. Integrating motion information over the whole image, however, is computationally expensive. This suggests representing the scene as a collection of simple shapes. Of all possible shape models, planes occupy a special place in that the projection of a plane undergoing rigid motion evolves according to a projective transformation. It is therefore natural to represent a scene as a collection of planes, which has been done often in the past, as for instance in [1, 17, 18, 20].

We seek to build on the strengths of direct methods, in order to avoid common problems with feature tracking by embedding the process in higher-level motion estimation, while keeping computational complexity at bay representing shape using a collection of simple templates.

1.3 Main contributions

We present an algorithm that integrates 2D region tracking and 3D motion estimation into a closed loop, therefore avoiding the local nature of point feature tracking. The input to the algorithm is a sequence of brightness images, and the output is the collective rigid motion of the scene.

Our algorithm integrates visual information in space as well as in time, building on the benefits of direct methods for SFM. Unlike most work in direct SFM, however, it relies on an explicit geometric and photometric model, providing a principled framework for detecting and rejecting outliers.

The computational model is causal and the algorithm recursive; its complexity makes it suitable for real-time implementation. The algorithm enjoys a number of analytical

properties, such as observability, which we prove in [8].

As a side benefit, our algorithm returns an estimate of the appearance of the scene as seen from an arbitrary pose, and could therefore be used for on-line constructions of 3D image mosaics. It can also be used for global alignment in long sequences, since the appearance of features once seen can be matched to current features in similar position and orientation. The template deformation models can be extended to take into explicit account changes in illumination or non-Lambertian reflection models [6].

2 From local photometry to global dynamics

Let S be a rigid, piecewise smooth surface in space, and $\mathbf{X} \in S$ the coordinates of a generic point on it. When seen from a moving frame, the coordinates change in time. If we let $\mathbf{X}_0 \doteq \mathbf{X}$, then we have $\mathbf{X}_t = R_t \mathbf{X}_0 + T_t$, where $R_t \in SO(3)^3$ and $T_t \in \mathbb{R}^3$ describe the rigid change of coordinates between the inertial (at time 0) and the moving frame (at time t). We assume to be able to measure, at each instant t , the irradiance $I(\mathbf{x}, t)$ at the point $\mathbf{x}_t = \pi(\mathbf{X}_t)^4$. We do not make distinction between the image coordinates and the homogeneous coordinates (with 1 appended). As a consequence of motion, the image undergoes a deformation that can be described by a nonlinear time-varying function of the surface S , $g_t^S(\cdot)$, such that

$$I(\mathbf{x}_0, 0) = I(g_t^S(\mathbf{x}_0), t) \quad (1)$$

when the surface is Lambertian. In general g is nonlinear and depends on an infinite number of parameters (a representation of the surface S):

$$g_t^S(\mathbf{x}_0) = \pi(R_t \mathbf{x}_0 \lambda + T_t) \text{ with } \lambda \mid \mathbf{x}_0 \lambda = \mathbf{X}_0 \in S. \quad (2)$$

However, one can restrict the class of functions g to depend upon a finite number of parameters (corresponding to a finite-dimensional parameterization of S), and therefore represent image deformations as a parametric class.

2.1 A generative model

There is a very simple instance when image deformations are captured by a finite-dimensional deformation model, that is when we restrict the class of surfaces to planes with unknown normal vector $\frac{\nu}{\|\nu\|} \in \mathcal{S}^2$ and intercept $\|\nu\|$. In fact, it is well known that a plane not passing through the origin (the optical center) can be described as $\Pi = \{\mathbf{X} \mid \nu^T \mathbf{X} = 1\}$, and therefore

$$g_t^\Pi(\mathbf{x}_0) = (R_t + T_t \nu^T) \mathbf{x}_0. \quad (3)$$

³ $SO(3)$ stands for the space of 3×3 rotation matrices (orthogonal with determinant 1).

⁴Where π denotes a camera projection, for instance, in the central projection case: $\pi(\mathbf{X}) = \frac{\mathbf{x}}{\|\mathbf{x}\|}$.

Given any matrix $M_t \in \mathbb{R}^{3 \times 3} / \mathbb{R}$ with rank at least 2, it can be shown [8] that it is in one to one correspondence with matrices of the form $R_t + T_t \nu^T$. Therefore, if the scene consists of a single planar surface, we can integrate photometric information on the entire surface by finding the matrix M that minimizes a discrepancy measure between $I(\mathbf{x}_0, 0)$ and $I(M_t \mathbf{x}_0, t)$ integrated over \mathbf{x}_0 ranging in the entire image domain D ; for instance

$$\hat{T}_t, \hat{R}_t, \hat{\nu} = \arg \min \int_D \|I(\mathbf{x}, 0) - I(M_t \mathbf{x}, t)\| d\mathbf{x} \quad (4)$$

for some choice of norm $\|\cdot\|$. Notice that the residual to be minimized is computed in the space of irradiance functions, and that the current model M_t , together with the first image $I(\mathbf{x}_0, 0)$, can be used to predict the next image $I(\mathbf{x}_{t+1}, t+1)$. In this sense this model is generative.

Of course, planes are quite a restrictive class of surfaces. However, we can use the above residual to test the hypothesis that a region of the image corresponds to (is well approximate by) a plane in space. Away from discontinuities, the larger the curvature, the smaller the region that will pass the test. By running the test all over the image (or on the portion of it that corresponds to high gradient of the irradiance, so as to eliminate at the outset regions with little or no texture), we can segment the image into a number of patches that correspond to planar approximations of the surface S . Obviously discontinuities and occluding boundaries will fail the test and therefore be rejected as outliers.

As a result of the procedure thus described, we are left with describing a surface with a certain number K of planar patches with normals ν^1, \dots, ν^K , all undergoing the same rigid motion T_t, R_t . Photometric information is integrated within each patch, while geometric and dynamic information is integrated across patches. In this sense, this model describes the scene using *local photometry and global dynamics*. A model of the time evolution of all the unknown quantities is therefore

$$\begin{cases} \nu_{t+1}^i = \nu_t^i & i = 1 \dots K \\ T_{t+1} = \exp(\hat{\omega}_t) T_t + V_t \\ R_{t+1} = \exp(\hat{\omega}_t) R_t \\ V_{t+1} = V_t + n_V(t) \\ \omega_{t+1} = \omega_t + n_\omega(t) \\ I(\mathbf{x}_0^i, 0) = I(\pi((R_t + T_t \nu_t^{i,T}) \mathbf{x}_0^i), t) + w_t \quad \forall \mathbf{x}_0^i \in D^i \end{cases} \quad (5)$$

where $n_V(t)$ denotes the unknown linear acceleration, $n_\omega(t)$ the rotational acceleration, and D^i is the region of the image that corresponds to the approximation of the surface S by the i -th planar patch with normal ν^i . The noise term w_t is modeled as an independent sequence identically distributed in such a way as to guarantee that the measured image I is positive.

3 Causal estimation

Having agreed to represent a surface as a rigid collection of planar patches supporting a radiance function that can deform according to a projective model, we can describe the unknown parameters (plane normals, rigid motion and velocity) as the state and input of a nonlinear dynamical system (5). Causally inferring a model of the scene then corresponds to estimating the state of the model (5) from its output (measured images). In order to arrive at a computationally simple solution to this problem, we will make a number of assumptions on the initial conditions and driving noises of the model (5).

3.1 Nonlinear filter and implementation

The first step towards implementation is to choose a local coordinate for model (5). To this end, we represent $SO(3)$ locally in canonical exponential coordinates: let Ω be a vector in \mathbb{R}^3 , then a rotation matrix can be represented by $\widehat{\Omega} \in so(3)$ such that $R = \exp(\widehat{\Omega})^5$. It is clear from the measurement equation in (5) that a scale factor between ν and T has to be fixed as they appear only as a product. Since we know that for a plane to be visible, the z component of its normal vector has to be strictly positive, we choose to fix the z component of one normal with a constant, say 1.

Once the model (5) is written in local coordinates it is immediate to use an extended Kalman filter to estimate the state. We follow the procedure described in [3]. For the filter to work in practice, one has to take into consideration of occlusions. During the camera motion, objects in the scene may occlude each others and hence cause some image patches to become not available. On the other hand, some new image patches can become visible. When a patch disappears (or is rejected by the hypothesis test described in section 3.3), we simply remove the corresponding normal vector from the state. When there a new patch appears, we first estimate its normal with a reduced filter and once its estimation is stable we insert it into the state.

3.2 Uniqueness of reconstruction

We are interested in reconstructing the state of the model (5), i.e. structure and motion of the scene, from the measurements on the image plane of a planar patch. It is natural to ask whether this reconstruction yields a unique solution or not. In system theory a necessary notion of uniqueness is captured by the concept of *observability*. In [8], we prove the following result:

⁵Rodrigues' formula is a convenient way to compute the exponential [16].

Proposition 1 *Given two planes with different normals in the scene, and a set of points in general configuration ⁶, there exists a unique reconstruction of the state from measurements based on model (5), if the translational velocity is non-zero.*

3.3 Outlier rejection

Since we are modeling local surfaces with planes, we have to detect regions where planes are not a good approximation for the surfaces in the scene. Since the measurements are obtained directly from the image plane, one has to consider outliers coming, for instance, from occluding boundaries and changes in illumination. The quality of a patch can be computed as the norm of the difference between the brightness of two image patches taken at two time instants, for instance the SSD (sum of squared difference [9]). We model the detection procedure as a hypothesis test. Let d_t^i be the sum of the squared intensity difference of patch i at time t , and σ^i be the variance of patch i at time 0.

$$\begin{aligned} H_0 : d_t^i &\leq \rho \sigma^i \\ H_1 : d_t^i &> \rho \sigma^i \end{aligned}$$

where $\rho > 0$.

Test H_0 says that a patch i is valid and H_1 says that it is not. The reason to use σ^i in a hypothesis test is that a patch with big variance in its initial appearance tends to have big difference. Therefore, there does not exist a single threshold for all d_t^i . However, if we take into account σ^i , we can find a single ρ for all the patches.

4 Experiments

4.1 Simulations with ground truth

Figure 1 shows the setting for the test on synthetic data. The scene consists in two planes in front of the viewer. They rotate along the axis parallel to both planes and passing through their centroid. The rotation is repeated six times and the corresponding camera motion (translation and rotation components) is shown in dashed lines in Figure 2. We assume no prior knowledge on the normals and initialize them with vectors aiming towards the viewer ($\nu = [0 \ 0 \ 1]^T$). The patches we choose are windows of 11×11 pixels. Since we have the ground truth, we can verify that the normals converge to the true value after about 500 frames. Figure 3 shows the estimated normals versus the ground truth. Figure 2 shows the estimated translation and rotation. Also in Figure 1 we visualize the estimated normals and their tangent planes for ease of visualization.

⁶We say that points in the image plane are in general configuration if there exist four points and such that none of any three among these four are collinear.

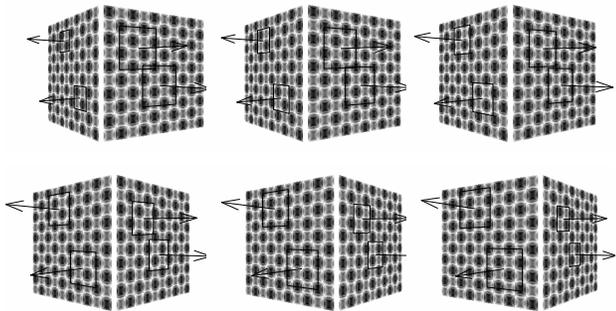


Figure 1. Synthetic dataset: the estimated normal vectors and their tangent planes are shown superimposed to six images from the original dataset. Estimation error is shown in Figure 3.

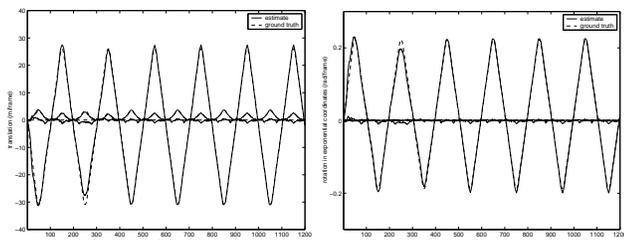


Figure 2. Estimated motion with ground truth: components of the translation vector (left) and exponential coordinates of rotation matrix (right).

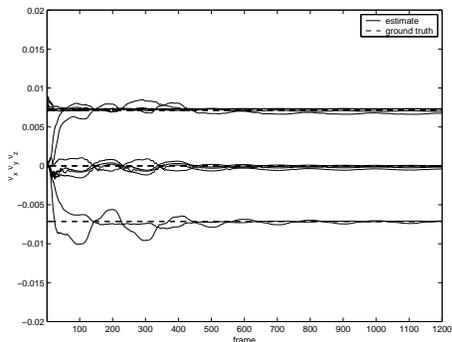


Figure 3. Estimated normals with ground truth.

4.2 Real images

Figure 5 is one image from a 700-frame long sequence. A handful of objects are on a chair that is rotated back and forth to the initial position. In Figure 5 we show four enlarged patches. They are the reconstruction of the corresponding patches at time 0, using the estimated camera motion and scene structure. Figure 6 shows the estimated camera motion.



Figure 4. Six images from the original sequence. The estimated normal vectors are shown superimposed to the real scene.

5 Discussion and future directions

We have presented a direct method to estimate motion and structure causally from image sequences. Instead of using point features we use planar patches, while non-planar patches and outliers are rejected using a simple hypothesis test. An extended Kalman filter is used to implement the algorithm in a causal fashion. The uniqueness of reconstruction of the filter has been proven in the attached technical report. We perform experiments on both synthetic and real scenes. Since we estimate motion as well as surface normals at the same time, we can explicitly take into consideration changes in illumination. We plan to incorporate changes of illumination explicitly within our framework. Even though the model we describe uses planes as primitives, the algorithm can be readily extended to any parametric representation of non-planar surfaces.

References

- [1] J. Alon and S. Sclaroff. Recursive estimation of motion and planar structure. In *IEEE Computer Vision and Pattern Recognition*, pages II:550–556, 2000.
- [2] A. Azarbayejani and A.P. Pentland. Recursive estimation of motion, structure, and focal length. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(6):562–575, June 1995.

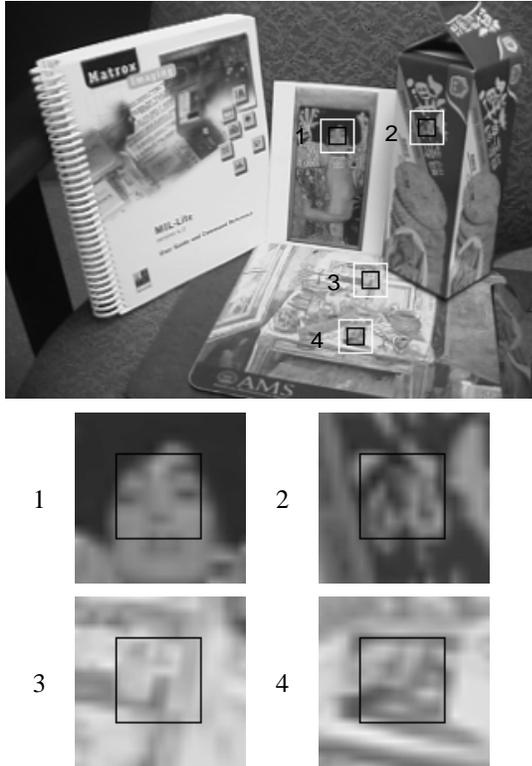


Figure 5. Visualization of the estimated albedo: estimated planar patches are seen superimposed to the corresponding view, with their normal vectors (top). The same estimated albedo is shown inside a black frame superimposed to the true albedo. As it can be seen, for the most part the alignment is perfect, although a slight misalignment can be noticed in patch 4. Having a sufficient number of patches would allow one to create an on-line image mosaic.

- [3] A. Chiuso, P. Favaro, H. Jin, and S. Soatto. “mfM”: 3-d motion from 2-d motion causally integrated over time: Implementation. In *European Conference on Computer Vision*, pages 735–750, 2000.
- [4] S. Christy and R. Horaud. Euclidean shape and motion from multiple perspective views by affine iterations. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(11):1098–1104, November 1996.
- [5] F. Dellaert, S. Seitz, C. Thorpe, and S. Thrun. Structure from Motion without Correspondence. *IEEE Computer Vision and Pattern Recognition*, 2:557–64, June 2000.
- [6] A.S. Georghiadis, D.J. Kriegman and P.N. Belhumeur. Illumination Cones for Recognition under Variable Lighting: Faces. In *IEEE Computer Vision and Pattern Recognition*, pages:52–59, 1998.
- [7] K.J. Hanna. Direct multi-resolution estimation of ego-motion and structure from motion. In *Workshop on Visual Motion*, pages 156–162, 1991.
- [8] H. Jin, P. Favaro and S. Soatto. Beyond Point Features, Integrating Photometry and Geometry in Space and Time: A Semi-direct Ap-

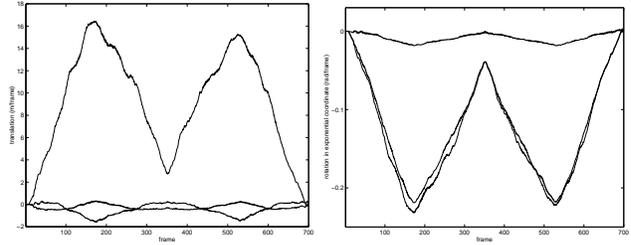


Figure 6. Estimated motion: (left) translation (right) rotation. Two cycles of rotation about the vertical axis were performed with return to the original pose. Although ground truth is not available, it can be seen that the estimated pose (rotation and translation) returns to zero as expected.

proach to Structure From Motion. *Technical report*, CSD-200034, UCLA, 2000.

- [9] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Image Understanding Workshop*, pages 121–130, 1981.
- [10] Y. Ma, J. Kosecka, and S. Sastry. Linear differential algorithm for motion recovery: A geometric approach. *International Journal of Computer Vision*, 36(1):71–89, January 2000.
- [11] L.H. Matthies, R. Szeliski, and T. Kanade. Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3(3):209–238, September 1989.
- [12] P.F. McLauchlan. A batch/recursive algorithm for 3d scene reconstruction. In *IEEE Computer Vision and Pattern Recognition*, pages II:738–743, 2000.
- [13] J. Rissanen. Information in prediction and estimation. In *CDC*, volume 1, pages 308–10, 1983.
- [14] H.S. Sawhney. Simplifying motion and structure analysis using planar parallax and image warping. In *International Conference on Pattern Recognition*, pages A:403–408, 1994.
- [15] L.S. Shapiro, A. Zisserman, and M. Brady. Motion from point matches using affine epipolar geometry. In *European Conference on Computer Vision*, pages B:73–84, 1994.
- [16] S. Soatto and P. Perona. Reducing structure-from-motion: A general framework for dynamic vision part I: Modeling. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(9):933–942, September 1998.
- [17] P. Sturm. Algorithms for plane-based pose estimation. In *IEEE Computer Vision and Pattern Recognition*, pages I:706–711, 2000.
- [18] P.F. Sturm and S.J. Maybank. A method for interactive 3d reconstruction of piecewise planar objects from single images. In *British Machine Vision Conference*, page Single View Techniques, 1999.
- [19] R. Szeliski and S.B. Kang. Direct methods for visual scene reconstruction. In *Representation of Visual Scenes*, pages 26–33, 1995.
- [20] R. Szeliski and P.H.S. Torr. Geometrically constrained structure from motion: Points on planes. In *3D Structure from Multiple Images of Large-Scale Environments*, pages 171–86, 1998.
- [21] J.I. Thomas and J. Oliensis. Recursive multi-frame structure from motion incorporating motion error. In *Image Understanding Workshop*, pages 507–513, 1992.
- [22] J. Weng, N. Ahuja, and T.S. Huang. Optimal motion and structure estimation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(9):864–884, September 1993.