International Journal of Computer Vision 52(1), 45–64, 2003

Constrained Structure and Motion From Multiple Uncalibrated Views of a Piecewise Planar Scene

ADRIEN BARTOLI AND PETER STURM
INRIA Rhône-Alpes, 655, avenue de l'Europe, 38334 Saint Ismier cedex, France
Adrien.Bartoli@inria.fr
Peter.Sturm@inria.fr

Received March 14, 2001; Revised September 13, 2002; Accepted September 13, 2002

Abstract. This paper is about multi-view modeling of a rigid scene. We merge the traditional approaches of reconstructing image-extractable features and of modeling via user-provided geometry. We use features to obtain a first guess for structure and motion, fit geometric primitives, correct the structure so that reconstructed features lie exactly on geometric primitives and optimize both structure and motion in a bundle adjustment manner while enforcing the underlying constraints. We specialize this general scheme to the point features and the plane geometric primitives. The underlying geometric relationships are described by multi-coplanarity constraints. We propose a minimal parameterization of the structure enforcing these constraints and use it to devise the corresponding maximum likelihood estimator. The recovered primitives are then textured from the input images. The result is an accurate and photorealistic model.

Experimental results using simulated data confirm that the accuracy of the model using the constrained methods is of clearly superior quality compared to that of traditional methods and that our approach performs better than existing ones, for various scene configurations. In addition, we observe that the method still performs better in a number of configurations when the observed surfaces are not exactly planar. We also validate our method using real images.

Keywords: 3D reconstruction, piecewise planar scene, constrained structure and motion, maximum likelihood estimator

1. Introduction

The general problem of scene modeling is, given a sequence of images without a priori information, to recover a model of the scene as well as (relative) pose and calibration. Performing this task accurately is one of the key goals in computer vision.

In this paper, we focus on the geometric scene modeling, i.e. we do not address aspects of lighting and surface appearance recovery besides perspective correction of texture maps. We aim at devising a framework for the recovery of photorealistic and accurate models from a sparse set of images.

Existing works fall into two categories: the *feature-* and the *primitive-based* approaches. By *features*, we designate two- or lower-dimensional geometric entities that can be extracted from individual images (e.g. points, lines, conics). By *primitives*, we mean other entities, e.g. planes or higher-dimensional ones (cubes, spheres). Let us examine these two approaches in more detail. First, the *primitive-based* approach, see e.g. Debevec et al. (1996), Lang and Förstner (1996), and Streilein and Hirschberg (1995), in which the user typically provides parametric primitives through a modeling program. Parameters are determined using 3D-2D or 2D-2D matches and possibly refined using photometric criteria, such as maximization of the gradient for wireframe models, to optimize their reprojection. If necessary, camera calibration is performed and texture maps are extracted for each primitive to produce a renderable model. This approach has proven to give convincing results, notably in terms of producing photorealistic rendering.

The *feature-based* approach, see e.g. Beardsley et al. (1996), relies on the existence of extractable image features. These features are matched accross the different views, typically using photometric and geometric criteria or by hand. From these, structure and motion are recovered. If necessary, camera calibration is performed and parameters refined in a bundle adjustment manner. This approach has proven to provide accurate reconstruction results, due to the high (in general) number of features considered. The problem is that modeling a scene with features alone does not allow to produce photorealistic rendering. Several works consider this issue, by using as features all the pixels, via dense matching (Pollefeys et al., 2000), space-carving (Kutulakos and Seitz, 1999; Seitz and Dyer, 1997), or plenoptic modeling (Gortler et al., 1996; Levoy and Hanrahan, 1996). The main limitation of at least the latter approach is that a high number of images is necessary to achieve accurate reconstruction. Other approaches relying on an a priori known environment (e.g. using turn-table sequences (Niem, 1994; Szeliski, 1993) or apparent contours (Cross and Zisserman, 2000)) can produce high quality rendering but do not work in the general case.

Actually, there exists a continuum between the two extreme feature- and primitive-based categories, made of *hybrid* approaches using both features and primitives.[1] These approaches are made to draw on the strength of both feature- and primitive-based categories: the high (in general) number of features might allow to obtain an accurate model recovery (even more accurate than for feature-based approaches) while the primitives contribute to form a photorealistic model. In hybrid approaches, the features and the primitives are linked by geometric constraints.

We study such an hybrid approach based on the point feature and the plane primitive. The geometric constraints used are incidence of points with none, one or several modeled planes and are called *multi-coplanarity constraints*. The corresponding constrained structure and motion recovery process is then called *piecewise planar structure and motion*.

These choices are motivated as follow. The point is a standard, widespread feature that may be easily extracted from the images. Most existing sparse structure and motion recovery systems deal with point features. The plane is a primitive sufficiently general to model a large number of real scenes, especially in man-made environments. Moreover, there are several works dealing with planes, that might be useful for an integrated modeling system: plane detection (Baillard and Zisserman, 1999; Berthilsson and Heyden, 1998; Dick et al., 2001; Faugeras and Lustman, 1988; Fornland and Schnörr, 1997; Sinclair and Blake, 1996; Tarel and Vézien, 1995), plane-guided point matching (Alon and Sclaroff, 2000; Faugeras and Lustman, 1988; Fornland and Schnörr, 1997; Sinclair and Blake, 1996; Viéville et al., 1995), and self-calibration using the knowledge of planes (Alon and Sclaroff, 2000; Malis and Cipolla, 2000; Triggs, 1998a; Viéville et al., 1995; Xu et al., 2000).

Concretely, we propose methods to parameterize points and planes under multi-coplanarity constraints. This parameterization is consistent in the sense that its number of parameters is the same as the number of degrees of freedom of the scene. It is employed to derive maximum likelihood estimators. Scene structure and camera motion are consistently estimated at once. A projective as well as a Euclidean version of the estimator are derived. The recovered structure perfectly satisfies the geometric constraints and is optimal in this respect, where optimal means maximum likelihood under a geometric error model.

In the following two sections, we present the piecewise planar structure and motion process and review existing work.

1.1. Piecewise Planar Structure and Motion

Given point correspondences between images, traditional *unconstrained structure and motion* reconstruct the points without using geometric constraints. First, suboptimal methods, see e.g. Beardsley et al. (1996) and Sturm and Triggs (1996), are used to compute an initial solution. The result is then refined using bundle adjustment (Slama, 1980; Triggs et al., 2000). If camera calibration is not available, the result is a projective reconstruction. In this case, the calibration information can be recovered online using several techniques (Heyden and Åström, 1998; Maybank and Faugeras, 1992; Pollefeys et al., 1998; Triggs, 1997). The uncalibrated reconstruction is then upgraded to metric and bundle adjustment is used to compute an optimal metric structure and motion.

In the projective case, when only points are used as features, then the scene has $11n - 15 + 3m$ essential degrees of freedom, where n is the number of views and m the number of points. Each view has 11 degrees of freedom; 15 degrees of freedom for the choice of the projective basis are deduced.

Assume now, that not only point correspondences are available, but also their plane memberships. The goal is to compute an optimal structure and motion including the geometric constraints underlying to the special multi-coplanar structure of the points. Ideally, this process is a maximum likelihood estimator optimizing features, primitive positions, and viewing parameters while enforcing the underlying geometric constraints. Consequently, there is a need for a new formulation of structure and motion, that models both features and primitives, and that preserves the relationships between them, in our case, that models points and enforces multi-coplanarity constraints. The use of such a constrained estimator has a strong impact on the structure and motion process. Compared to the unconstrained case, the use of primitives constituting an important geometric constraint on both structure and motion, we can expect better reconstruction results. It might also be faster, as the number of parameters is usually lower.

Intra-primitive constraints, such as a priori known angles or parallelism of the modeled planes could be used. One problem with these constraints is that, generally speaking, they can not be used in a projective framework. Many other kinds of constraints could be modeled, such as the collinearity of points.

Choosing the constraints to model is difficult. Indeed, this is a trade-off between accuracy (the more constraints are used, the more accurate the reconstruction will be) and the complexity of the algebraic modeling. If too many kinds of constraints are used, then we end up with a network of constraints, that may be viewed as a graph linking features and primitives, and that might be redundent in the sense of cycles in this graph. Another issue is the automatization of an integrated modeling system. High-order constraints, such as the arrangement of planes in e.g. cuboïds, are more difficult to detect than the coplanarity of a set of points. A comprehensive treatment of the possible geometric constraints is out of the scope of this paper.

As said before, the incorporation of multi-coplanarity constraints has an impact on the number of essential degrees of freedom of the scene, e.g. a point on one plane has 2 degrees of freedom instead of 3 in the unconstrained case. Consequently, the number of degrees of freedom of such a scene becomes equal to $11n - 15 + 3p + 3m - \sum_j jm_j$ where the notation m_j designates the number of points lying on j of a total of p modeled planes.

Let us review existing piecewise planar structure and motion estimators.

1.2. Previous Work

Most of the existing works yield only a sub-optimal estimation of the geometry. Actually, they fall into two categories:

- The recovered structure is only approximately piecewise planar so clearly the results can not be optimal (Faugeras and Lustman, 1988; McGlone, 1996; Szeliski and Torr, 1998; Tarel and Vézien, 1995; Xu et al., 2000);
- The recovered structure is piecewise planar but the method is not optimal because it can not be turned into a maximum likelihood estimator or only the single-coplanarity constraint is modeled (Alon and Sclaroff, 2000; Baillard and Zisserman, 1999; Bartoli et al., 2001).

If we want our estimator to be optimal with respect to piecewise planarity, it has to fall into the second category, i.e. the recovered model has to be exactly piecewise planar. The constrained structure and motion is a maximum likelihood estimator that incorporates points, planes and multi-coplanarity constraints in a bundle adjustment manner. The cost function is nonlinear (Slama, 1980; Triggs et al., 2000) and subject to constraints. There are several ways to conduct such an optimization process, in particular, we could use constrained optimization techniques such as sequential quadratic programming or a specific structure and motion parameterization enforcing the multi-coplanarity constraints (Bartoli and Sturm, 2001; Bartoli et al., 2001).

Ideally, these two possibilities give the same results because they are both consistent (i.e. the number of algebraic degrees of freedom is the same as that of essential degrees of freedom of the scene) and the cost function being optimized is the same. However, in practice, the convergence of the optimization process is determined by the number of parameters used which directly influences numerical stability. This determines which method to use in which case.

In our case, the number of parameters is high and so we have to reduce it to or close to the minimum, i.e. the number of essential degrees of freedom, if we want to ensure a stable optimization process. The first possibility consists in systematically adding parameters to the system to model constraints and is consequently unadapted. The second possibility is less systematic, so needs more algebraic manipulations to be derived. However, the number of parameters is so reduced that the convergence might be faster and more reliable. Another issue that is important to be dealt with for both estimation cost and stability is that of analytic differentiation for the non-linear minimizer, which implies that the parameterization has a closed-form expression.

We addressed the case of two views and points lying on one plane (i.e. the single-coplanarity constraint) in Bartoli et al. (2001) and extended it to multi-view and multi-coplanarity constraints in Bartoli and Sturm (2001) where we derived the maximum likelihood estimator but without the possibility of analytic differentiation. In this paper, we present an estimator and the corresponding parameterization which is minimal for the structure and quasi-minimal for the motion, for n views and a quasi-general set of multi-coplanarity constraints and which allows analytic differentiation.

As real world surfaces are only approximately planar, we experimentally evaluate the performance of the constrained method compared to an unconstrained one with respect to different degrees of deviation from planarity and different scene configurations.

Since our approach needs to upgrade an uncalibrated reconstruction to metric, we perform self-calibration. To initialize a bundle adjustment procedure, we use the linear method of Pollefeys (1999), inspired by Triggs (1997), for the estimation of variable focal length. In practice, we encountered a singular situation, that is likely to occur in modeling applications: the optical axes of all images meet in a single 3D point (which will usually be the center of the modeled object). We adapt the basic method to this case and validate the approach on real images.

In Section 2 we give our notations. We then present our parameterization and the corresponding maximum likelihood estimator for a projective framework in Section 3, followed by an equivalent scheme in the Euclidean case in Section 4 where we also present self-calibration. We report on experiments on simulated data for constrained structure and motion in Section 5. Finally, Section 6 shows experimental results obtained using real images taken with an uncalibrated camera which validate both the reconstruction and the self-calibration processes, followed by our conclusions.

2. Notations

Physical entities (points, planes, etc.) are typeset using italic fonts (X, π, etc.) and their corresponding homogeneous coordinate vectors using the same letters in bold fonts ($\mathbf{X}$, $\boldsymbol{\pi}$, etc.). Matrices are designated by sans-serif fonts such as H. Vector and matrix elements are typeset using italic fonts and indices, e.g. $\mathbf{X} \sim (X_1, X_2, X_3, X_4)^\top$ where $^\top$ is the transposition and $\sim$ the equality up to a non-zero scale factor.

The notation $\mathbf{X}_{/j}$ is used to designate the vector formed with the elements of $\mathbf{X}$ with index different from j. Similarly, $\mathbf{X}_{j \leftarrow \alpha}$ represents the vector $\mathbf{X}$ with the value α inserted at the j-th position. The cross product is written $\times$ and the associated 3×3 skew-symmetric matrix $[\cdot]_\times$.

We model cameras using perspective projection, described by a 3×4 homogeneous matrix P. Non-linear optimization processes are conducted using the Levenberg-Marquardt algorithm (Gill et al., 1981).

3. Constrained Projective Structure and Motion

In this section, we describe how to minimally parameterize the structure and quasi-minimally the motion in the projective case. We then derive the maximum likelihood estimator corresponding to the constrained structure and motion.

As shown in the expression for the number of essential degrees of freedom of the scene, we have to take into account 15 degrees of gauge freedom left by the arbitrary choice for the projective basis of the reconstruction. Gauge freedom is defined as the internal freedom of choice for a coordinate system (Triggs, 1998b). It can be fixed using a particular formulation for the structure (Heyden and Åström, 1995) or for the camera matrices (Beardsley et al., 1996). Due to the complexity of structure parameterization, we have chosen to absorb the gauge freedom into the parameterization of motion.

In the next two sections, we describe respectively our structure and motion parameterizations.

3.1. Structure Parameterization

As said in the introduction, we have to parameterize both planes and points and in addition enforce the underlying multi-coplanarity constraints. The parameterization consists in passing from the usual homogeneous 4-vectors that represent points and planes in

3D projective space, to a minimal set of parameters representing the structure while enforcing the multi-coplanarity constraints. We first give an homogeneous and consistent parameterization for planes and points and then remove the homogeneity to reach a minimal parameterization. This last step is achieved using what we call *mapped coordinates* that allow to locally remove homogeneity. This is also used in the parameterization of the motion and in the Euclidean case.

3.1.1. Multi-coplanarity Constraints. A multi-coplanarity constraint is a geometric constraint between a point and a set of planes. Before parameterizing the structure, we have to decide where, in the parameterization of planes, of points or both, these constraints have to be incorporated. Actually, it seems inevitable to incorporate them in the point parameterization. Let us investigate the case of plane parameterization. Indeed, consider the case of a point lying on more than three planes. Such a point does not have, in general, any degree of freedom, and can be determined using three of the planes it lies on.[2] Once this point has been determined, it constrains the position of the other planes. Consequently, plane parameterization is dependent on multi-coplanarity constraints provided they contain a point lying on more than three planes.

If we do not model points lying on more than three planes (or take into account only three of the planes they lie on), it is possible to parameterize each plane independently while the multi-coplanarity constraints up to three planes are to be taken into account only for point parameterization. Considering that points lying on four or more planes are rare, we make such an assumption (an algebraic solution will just be sketched). Let us see the corresponding parameterization.

3.1.2. Planes. As said above, planes do not incorporate multi-coplanarity constraints and each one has therefore 3 degrees of freedom. An homogeneous 4-vector is then a consistent parameterization.

3.1.3. Points Under Multi-coplanarity Constraints. We describe point parameterization performed under a local incorporation of multi-coplanarity constraints. Let us examine different possible means. We then present our solution for the different multi-coplanarity cases.

To simplify the reading, we consider the case of a 2D point x constrained to lie on a 2D line l, which is similar to the 3D single-coplanarity case. Such a constraint is expressed as $\mathbf{l}^\top \mathbf{x} = 0$ and is satisfied for any point expressed in the nullspace of $\mathbf{l}^\top \sim (l_1, l_2, l_3)$.

The approach that naturally comes to mind is to compute a basis for the nullspace of $\mathbf{l}^\top$ and to express the coordinates of point x in this basis. We examine two ways to compute this nullspace basis and show that each of them is not appropriate to our problem.

A basis for the nullspace of $\mathbf{l}^\top$ is given by the skew-symmetric 3×3 cross-product matrix associated to $\mathbf{l}$ (there are other possible bases):

$$\mathsf{L} \sim [\mathbf{l}]_\times \sim \begin{pmatrix} 0 & -l_3 & l_2 \\ l_3 & 0 & -l_1 \\ -l_2 & l_1 & 0 \end{pmatrix}.$$

One can easily check that, as required, $\mathbf{l}^\top \mathsf{L} = \mathbf{0}_3^\top$ and $\operatorname{rank} \mathsf{L} = 2$. Any point on l can be represented by a linear combination of the 3 columns of L, thereby involving 3 homogeneous coefficients. This is not consistent since a point on a line has only 1 degree of freedom. On the other hand, one could think of using only 2 columns of L as a basis for the nullspace, say drop the leading column $\mathbf{l}_1$. In this case, the representation is consistent, but it is no more complete: the point with coordinate $\mathbf{l}_1$ lying on l can not be represented as a linear combination of the two last columns of L.

Another possibility is to compute an orthonormal basis for the nullspace of $\mathbf{l}^\top$ through e.g. its singular value decomposition:

$$\mathbf{l}^\top \sim \mathbf{l}^\top \operatorname{diag}(1, 0, 0)(\mathbf{l}_{3\times 1} \,|\, \bar{\mathsf{V}}_{3\times 2}).$$

In this case, the basis given by the two columns of $\bar{\mathsf{V}}$ is minimal and the corresponding parameterization would be consistent. However, since the entries of $\bar{\mathsf{V}}$ do not depend directly on the coefficients of $\mathbf{l}$, analytic differenciation would not be possible in the underlying optimization process.

Consistency and analytic differenciation are the main reasons for our specific parameterization to be used. We successively deal with points lying on none, one, two and three planes.

3.1.3.1. Unconstrained Points. Such a point does not lie on any modeled plane and being therefore not subject to any modeled geometric constraint, it has 3 degrees of freedom. An homogeneous 4-vector is then a consistent parameterization.

3.1.3.2. Single-Coplanar Points. Let X be a point constrained to lie on a plane π. Such a point has 2 degrees of freedom and our goal is then to express it via an homogeneous 3-vector—instead of the general homogeneous 4-vector—by incorporating the single-coplanarity constraint.

Algebraically, this constraint is written as $\boldsymbol{\pi}^\top \mathbf{X} = 0$. Let us find a change of projective basis where each point lying on π has an element fixed to a constant value, so that this element can be ignored in the parameterization of X. For that purpose, we define the class of homographies H_π^j by the identity matrix of size 4 × 4 where the jth row ($j \in \{1, \ldots, 4\}$) has been replaced by the 4-vector $\boldsymbol{\pi}^\top$ (e.g. $\mathsf{H}_\pi^1 \sim (\begin{smallmatrix} & \boldsymbol{\pi}^\top & \\ \mathbf{0}_3 & & \mathrm{I}_{3\times 3} \end{smallmatrix})$). Let $\Xi \sim \mathsf{H}_\pi^j \mathbf{X}$ be the representation of X in this new basis. By definition of H_π^j ,we have $\Xi_j = 0$ and the point X can therefore be parameterized by $\Xi_{/j}$, the homogeneous 3-vector formed from the 3 elements of Ξ with index different from j, $\mathbf{X}$ being further recovered using $\mathbf{X} \sim (\mathsf{H}_\pi^j)^{-1}\,\Xi$.

There are 4 possibilities for the choice of j. Since $(\mathsf{H}_\pi^j)^{-1}$ is necessary to recover $\mathbf{X}$ from Ξ, we choose j as the index that maximizes (in magnitude) the determinant of H_π^j: $j = \mathrm{argmax}_i\, |\det \mathsf{H}_\pi^i|$ which in fact leads to $j = \mathrm{argmax}_i\, |\pi_i|$. Such a choice ensures H_π^j to be a bijective transformation since $\det \mathsf{H}_\pi^j = \pi_j$ that, by construction, is always non-zero. Indeed, $\boldsymbol{\pi}$ is an homogeneous vector and has therefore at least one non-zero element.

Table 1 shows the practical algorithm for parameterizing/unparameterizing $X \in \pi$ derived from the above reasoning. In the unparameterization, we divide by π_j that, as said above, is always non-zero.

The dropped coordinate depends on the current estimate of π. Therefore, it might change between two steps of the optimization process. However, this does not create discontinuities since after each optimization step, the structure is unparameterized and standard homogeneous coordinates are recovered. The structure is then reparameterized for the next iteration, and the index of the dropped coordinate may change. The parameterization is therefore used in a local manner, which is important in order to keep smooth the cost function.

Table 1. Parameterization/unparameterization of a single-coplanar point.

Let X be a point subject to a single-coplanarity constraint with plane π. The homogeneous 4-vector $\mathbf{X}$ represents X in the current projective basis while the homogeneous 3-vector $\tilde{\mathbf{X}}$ is a parameterization of X incorporating the single-coplanarity constraint.

Parameterization ($\mathbf{X} \rightarrow \tilde{\mathbf{X}}$):

- Choose j such that $j = \arg\max_i |\pi_i|$subject to $j \in \{1, \ldots, 4\}$ in the projective case and $j \in \{1, \ldots, 3\}$ in the Euclidean case;
- $\tilde{\mathbf{X}} \sim \mathbf{X}_{/j}$.

Unparameterization ($\tilde{\mathbf{X}} \rightarrow \mathbf{X}$):

- Compute $\alpha = -\frac{\sum_{i \neq j} \pi_i X_i}{\pi_j}$;
- $\mathbf{X} \sim \tilde{\mathbf{X}}_{j \leftarrow \alpha}$.

3.1.3.3. Multi-coplanar Points, Two Planes. Let X be a point constrained to lie on planes π and π'. Such a point has 1 degree of freedom provided that $\pi \neq \pi'$ and our goal is then to express it via an homogeneous 2-vector—instead of the general homogeneous 4-vector—by incorporating the multi-coplanarity constraint.

We follow the same reasoning as for the previous case. We define the class of homographies $\mathsf{H}_{\pi,\pi'}^{j,j'}$ by the matrix H_π^j where the j'-th row has been replaced by the 4-vector $\boldsymbol{\pi}'^\top$ (e.g. $\mathsf{H}_{\pi,\pi'}^{1,2} \sim (\begin{smallmatrix} & \boldsymbol{\pi}^\top & \\ & \boldsymbol{\pi}'^\top & \\ 0_{2\times 2} & & \mathrm{I}_{2\times 2} \end{smallmatrix})$). Let us consider $\Xi \sim \mathsf{H}_{\pi,\pi'}^{j,j'} \mathbf{X}$. By definition of $\mathsf{H}_{\pi,\pi'}^{j,j'}$, we have $\Xi_j = \Xi_{j'} = 0$ and point X can therefore be parameterized by $\Xi_{/j,j'}$, the homogeneous 2-vector formed from the 2 elements of Ξ with index different from j and j', $\mathbf{X}$ being further recovered using $\mathbf{X} \sim (\mathsf{H}_{\pi,\pi'}^{j,j'})^{-1}\,\Xi$.

Since j and j' must be different, this leaves 4 × 3 = 12 different choices for them. As $(\mathsf{H}_{\pi,\pi'}^{j,j'})^{-1}$ is needed, we choose j and j' such that the determinant of $\mathsf{H}_{\pi,\pi'}^{j,j'}$ is maximized (in magnitude). Subsequently deriving a practical algorithm as in the single-coplanarity case is then straightforward.

3.1.3.4. Multi-coplanar Points, Three Planes. Let X be a point constrained to lie on planes π, π' and π''. As already mentioned previously, it is straightforward to see that a point lying on three planes does not have, in general, any degree of freedom.[2] Such points are therefore not represented in the parameterization and have to be recovered from the three plane equations. There are two ways to do that. One can either choose a scheme similar to the one given previously or use the generalized cross-product, which gives a closed-form expression for the point (each point coordinate is given by the determinant of a 3 × 3 matrix of plane coefficients).

3.1.3.5. Multi-coplanar Points, More Than Three Planes. As said previously, this case is rare. Dealing with it properly would add a great complexity to the system, in the sense that constraints would then be expressed not only on points but also on planes, thereby creating a graph of constraints with possible redundancies and cycles. Let us sketch, however, how the case of a point X lying on 4 planes π, π', π'' and π''' could be handled algebraically. Other higher order cases, though more complicated, could then be handled in a similar manner. The constraints are express as:

$$\mathsf{B}^\top \mathbf{X} = \mathbf{0}_4 \quad \text{where } \mathsf{B}_{4\times 4} \sim (\boldsymbol{\pi} \;\; \boldsymbol{\pi}' \;\; \boldsymbol{\pi}'' \;\; \boldsymbol{\pi}''').$$

This equation means that the matrix B has a (at least) 1-dimensional nullspace, i.e. det $\mathsf{B} = 0$, which yields a 4-linear constraint on the coefficients of the plane equations. If one chooses to constrain e.g. plane π, then one of its coordinates may be dropped by considering the above-derived equation, and by applying a scheme similar to that described in Table 1, for the single-coplanarity case.

3.1.3.6. Modeling Intra-primitive Metric Constraints. In this paragraph, we give some hints on the algebraic modeling of intra-primitive constraints, and in particular on the perpendicularity and the orthogonality of planes. As explained in the introduction, a comprehensive treatment of all these constraints is out of the scope of this paper.

Firstly, consider the perpendicularity of two planes π and π'. This constraint can be algebraically expressed by considering that the dot product between the normal vectors of two such planes must vanish:

$$\pi_1\pi_1' + \pi_2\pi_2' + \pi_3\pi_3' = 0.$$

This bilinear constraint can be enforced by the elimination of one parameter to contrain one of the two planes to be perpendicular to the other one. We end up with the same problem as that of modeling the single-coplanarity constraint described above.

Secondly, consider the modeling of the parallelism of two planes π and π'. The normal vectors of two such planes must be equal, up to scale, which is equivalent to nullifying there cross-product:

$$\begin{cases} \pi_2\pi_3' - \pi_3\pi_2' = 0 \\ \pi_3\pi_1' - \pi_1\pi_3' = 0 \\ \pi_1\pi_3' - \pi_3\pi_1' = 0. \end{cases}$$

Among these 3 equations, only 2 are independent, but one can not choose 2 of them a priori. Therefore, depending of which plane is to be contrained and on which axes, 2 equations are used to eliminate 2 of its parameters. Since these equations are bilinear, we end up with the same problem as that of modeling the multi-coplanarity constraint with 2 planes, described previously.

3.1.4. Mapped Coordinates. Homogeneous algebraic entities have an internal gauge freedom as they are only defined up to a non-zero scale factor. Consequently, they are not minimal in the sense that they are overparameterized. We define a tool called *mapped coordinates* that locally removes the homogeneity, in other words produces a minimal version of an homogeneous entity. Let us consider the case of homogeneous vectors of $\mathbb{P}^\nu$, which is not a restriction, the method being valid for any homogeneous entity (matrix, tensor). In more detail, a $(\nu + 1)$-vector $\mathbf{v}$, can be decomposed into a ν-vector $\tilde{\mathbf{v}}$ and a map $\mu \in \{1, \ldots, \nu + 1\}$, the index of a coefficient to be fixed. An important property is that slightly changing $\mathbf{v}$ does not, in general, affect μ but only $\tilde{\mathbf{v}}$, and if μ is affected, it will usually not create numerical instability (in the sense that the maximum coefficient of $\mathbf{v}$ will not tend towards zero during e.g. optimization).

The map μ is chosen as the index of the entry of $\mathbf{v}$ that has the largest absolute value. This choice can be justified as follows. If we assume that all entries of $\mathbf{v}$ have the same probability to become zero during an optimization step, our choice minimizes the probability that the selected entry (i.e. the one corresponding to the map μ) vanishes.

Consequently, this system is adapted to non-linear optimizers such as Levenberg-Marquardt (Gill et al., 1981), where the map can be re-estimated at each step of the optimization process. A practical algorithm for using mapped coordinates is given in Table 2.

Table 2. Mapped coordinates for homogeneous entities. Only $\tilde{\mathbf{v}}$ has to be included in optimization processes.

Let $\mathbf{v}$ be an homogeneous $(\nu + 1)$-vector. Any other homogeneous entity (matrix, tensor) can be brought back to this case by stacking its elements into a single vector. The inhomogeneous ν-vector $\tilde{\mathbf{v}}$ represents the mapped coordinates of $\mathbf{v}$ whereas the integer μ represents its map.

Mapping $(\mathbf{v} \rightarrow (\tilde{\mathbf{v}}, \mu))$:

- Choose μ such that $\mu = \arg\max_i |\mathbf{v}_i|$;
- $\tilde{\mathbf{v}} = \frac{\mathbf{v}/\mu}{v_\mu}$.

Unmapping $((\tilde{\mathbf{v}}, \mu) \rightarrow \mathbf{v})$:

- $\mathbf{v} \sim \tilde{\mathbf{v}}_{\mu \leftarrow 1}$.

3.1.5. Summary of Structure Parameterization. We have given algorithms to exploit multi-coplanarity constraints for up to three planes per point. These constraints are enforced in an homogeneous manner while reducing the number of parameters for each point, see Section 3.1.3, and the homogeneity is removed using mapped coordinates, as indicated in Table 2, to obtain a minimal parameterization.

3.2. Motion Parameterization

In this section, we first parameterize camera projection matrices in an homogeneous manner and then remove the homogeneity using mapped coordinates to obtain a quasi-minimal parameterization.

We have chosen previously to fix the projective reconstruction basis via the camera parameterization. It has then to express $11n - 15$ degrees of freedom but actually has $10 + 11(n - 2)$ parameters (see below), i.e. is overparameterized by 3. This is not a problem for the optimization process since this number does not depend neither on the number of views nor on the number of points.

The number of parameters is obtained as follows. Each of the n views is represented by 11 parameters from its camera matrix, except for 2 of them, related by the epipolar geometry (or equivalently, one special-form projection matrix), that we represent using 10 parameters. More details are given below, where we describe the geometry of one, two, three or more views. Note that the motion parameterization is independent from the structure, and in particular, does not depend on the fact that the structure is constrained or not.

One View. The projective reconstruction basis can not be uniquely fixed. However the camera matrix P can be arbitrarily set, e.g. we use here $\mathsf{P} \sim (\mathrm{I} \mid \mathbf{0})$.

Two Views. If we suppose that the first camera matrix has been fixed, the second one has 7 degrees of freedom. Indeed, the geometry of such a system is described by the epipolar geometry contained in the rank deficient fundamental matrix F. Provided P has the form given above, the second camera matrix can be extracted from F as $\mathsf{P}' \sim ([\mathbf{e}']_\times \mathsf{F} \mid \mathbf{e}')$ where $\mathbf{e}'$ is the second epipole defined by $\mathsf{F}^\top \mathbf{e}' = 0$.

Minimally parameterizing the rank-2-ness of the fundamental matrix requires the use of several maps (Bartoli et al., 2001; Zhang, 1998) which is complicated from an implementation point of view. Alternatively, it is possible to overparameterize rank-2-ness by using a plane homography H and the second epipole e'. The second camera matrix is then formed as $\mathsf{P}' \sim ([\mathbf{e}']_\times^2 \mathsf{H} \mid \mathbf{e}')$ where $[\mathbf{e}']_\times^2 \mathsf{H}$ is the *canonical plane homography* which is the only plane homography satisfying $\mathsf{H}^\top \mathbf{e}' = 0$ (Bartoli and Sturm, 2001) (it is thus singular).

In this paper, we use this second possibility. The problem is parameterized by the 8 mapped coordinates of H and the 2 mapped coordinates of e', which yield 10 parameters. Consequently, it is overparameterized by $10 - 7 = 3$ parameters, since the two-view motion has only 7 degrees of freedom.

Three or More Views. Two or more views completely fix the projective basis. Consequently, each additional view adds 11 degrees of freedom to the system and in the general case their camera matrices do not have any special form and have to be entirely parameterized. We use mapped coordinates for that purpose.

The motion parameterization is summarized in Table 3.

3.3. Maximum Likelihood Estimator

We describe the maximum likelihood estimator for constrained structure and motion using the previously described parameterization. We first analyze which kinds of points are reconstructable and under which conditions, notably if they have to be included in the constrained optimization process. We then show how to initialize the parameterization from a general structure

Table 3. Motion parameterization. Notations $\tilde{\mathsf{H}}$, $\tilde{\mathbf{e}}'$ and $\tilde{\mathsf{P}}_k$ respectively designate the mapped coordinates (see Table 2) of the canonic plane homography (see text), of the second epipole (i.e. the projection of the first camera's center of projection onto the image plane of the second camera) and of other camera matrices. dof stands for degrees of freedom.

No. of views	No. of dof	No. of parameterization	Parameters	Gauge constraints
$n = 2$	7	10	$\tilde{\mathsf{H}}, \tilde{\mathbf{e}}'$	$\mathsf{H}^\top \mathbf{e}' = \mathbf{0}$
$n \geq 3$	$7 + 11(n - 2)$	$10 + 11(n - 2)$	$\tilde{\mathsf{H}}, \tilde{\mathbf{e}}', \tilde{\mathsf{P}}_{k \geq 3}$	$\mathsf{H}^\top \mathbf{e}' = \mathbf{0}$

and motion (when multi-coplanarity constraints are not enforced), in the case of motion and then structure. Finally, we give the cost function and details on the maximum likelihood estimator.

3.3.1. Initialization. At this step, we suppose to have a first guess of structure and motion as well as a clustering of points into multi-coplanar groups, see Section 6.

Feature Reconstructability. Planes are reconstructable provided that at least three points that they contain can be themselves reconstructed without geometric constraints. Once planes are reconstructed, new point reconstructions can be obtained. Table 4 gives which points, in terms of the number of views they are seen in and number of planes they lie on, can be reconstructed and if they have to be incorporated in the optimization process (i.e. if they add redundancy useful for optimization).

Motion. We have to change the projective basis such that the first camera matrix becomes $(\mathrm{I} \mid \mathbf{0})$. This is done by post-multiplying all camera matrices by an appropriately chosen 3D homography and pre-multiplying the structure by the inverse of this homography.

Constrained Structure. The initialization of points depending on that of planes, we first estimate plane equations and then points.

A plane is fitted to the points of each coplanar group. If X is a point lying on the plane π, the constraint $\mathbf{X}^\top \boldsymbol{\pi} = 0$ holds. By stacking the equations for all points lying on the plane, we obtain a linear system for $\boldsymbol{\pi}$ which can be solved using e.g. singular value decomposition. Another possibility is to estimate a plane homography between two images of the plane and to further extract the plane equation.

Table 4. Summary of which points are reconstructable under which condition. "unconstrained" indicates a reconstruction when planes are not yet modeled, "optimization" indicates a reconstruction possible using planes and for points that add redundancy useful for optimization and "constrained" indicates a reconstruction possible only after the maximum likelihood estimation.

No. of views	No. of planes	Unconstrained	Optimization	Constrained
0	0	No	No	No
	1	No	No	No
	2	No	No	No
	≥3	No	No	Yes
1	0	No	No	No
	1	No	No	Yes
	≥2	No	Yes	No
≥2	Any	Yes	Yes	No

The unconstrained points and the multi-coplanar points lying on three or more planes are easy to initialize. Indeed, the former are not subject to any modeled geometric constraint and are taken directly from the initial structure, and the latter do not have any degree of freedom and so do not need initial values.

On the other hand, single-coplanar and multi-coplanar points lying on two planes need a special initialization. As we work in projective space, we can not consider any metric in space (such as orthogonal projection) and have to do measurements in the images.

For a single-coplanar point X lying on a plane π, we consider one of its projections and reconstruct the 3D point by intersecting the associated viewing ray with the plane π. We measure the reprojection error in all images where X is visible. We iterate over the set of images where X is visible and select the one that minimizes the total reprojection error.

For a multi-coplanar point X lying on planes π and π', we adopt the same method. However, to ensure that the reconstructed point lies on the two planes, we orthogonally project one of its image points onto the projection of the intersection line of π and π' and then reconstruct as above. Which plane π or π' is used to reconstruct does not matter. Details for this initialization are given in Bartoli and Sturm (2001).

3.3.2. Optimization. Our goal is to derive an optimal estimator, in the sense of the maximum likelihood, for points and planes under multi-coplanarity constraints. This result is obtained by enforcing exactly the constraints, as is does by our parameterization. The cost function to minimize is the root mean square or, equivalently, the sum of square of the reprojection residuals (Slama, 1980; Triggs et al., 2000). In fact, this gives the maximum likelihood estimator under the assumption that errors in image point positions are identically and independently distributed according to a centered Gaussian, or normal, law.

We also include camera motion parameters into the non-linear optimization since an independent computation of the maximum likelihood estimate for the structure only is not possible: both structure and motion have to be estimated at once.

The cost function, denoted by $\mathcal{C}$, depends on measured image points x_{ij} and on reprojected points $\hat{x}_{ij}$ predicted by structure and motion parameters $\mathcal{S}$. It is defined by:

$$\mathcal{C}(\mathcal{S}) = \sum_i \sum_j w_{ij}\, d^2(\mathbf{x}_{ij}, \hat{\mathbf{x}}_{ij}).$$

Indices i and j respectively represent the different images and the different structure points and $d(.,.)$ is the Euclidean distance. We set $w_{ij} = 1$ if and only if the j-th point appears in the i-th image and 0 otherwise. The optimal structure and motion parameters $\hat{\mathcal{S}}$ are then given by the minimization of $\mathcal{C}$ over $\mathcal{S}$:

$$\hat{\mathcal{S}} = \arg\min_{\mathcal{S}} \mathcal{C}(\mathcal{S}).$$

This is done in practice using the Levenberg-Marquardt algorithm with analytic differentiation.

Let us investigate how to upgrade the obtained structure and motion to a metric frame.

4. Constrained Euclidean Structure and Motion

In this section, we describe how to upgrade the previously recovered projective structure and motion to metric and how to parameterize them in order to obtain a constrained maximum likelihood estimator.

4.1. Upgrade to Metric

There exist several possibilities to upgrade a projective reconstruction to metric, without a full prior calibration, e.g. by providing constraints on scene structure, camera motion, or calibration. In this work, we perform self-calibration. A Euclidean bundle adjustment is initialized using the linear method of Pollefeys (1999), inspired by Triggs (1997), that assumes known intrinsic parameters, besides the variable focal length. The method is rather straightforward, but we describe it here since the basic method is subject to a degenerate situation we encountered in practice, and that is likely to occur quite often in modeling applications for e.g. built environments. We give a variant of the method that does not degenerate in this case.

Suppose that P_i are the projection matrices associated with the projective reconstruction obtained so far. We suppose that all the intrinsic parameters are given, besides the focal lengths, f_i, for the individual images. In practice, we assume the principal points (u_i, v_i) to lie in the center of the respective image, and we know the cameras' aspect ratios τ_i (in fact, they could easily be included in the linear self-calibration routine). The skew factor is neglected, i.e. we assume pixels to be rectangular (in the linear method; skew is estimated during bundle adjustment).

Self-calibration is based on estimating a projective transformation T such that the transformed projection matrices can be decomposed into extrinsic and intrinsic parameters, such that the latter have the known values, i.e.:

$$\exists f_i, \mathsf{R}_i, \mathbf{t}_i \colon \mathsf{P}_i\mathsf{T} \sim \begin{pmatrix} \tau_i & 0 & u_i \\ 0 & 1 & v_i \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} f_i & 0 & 0 \\ 0 & f_i & 0 \\ 0 & 0 & 1 \end{pmatrix} (\mathsf{R}_i \mid \mathbf{t}_i),$$

where the R_i are orthonormal matrices and the $\mathbf{t}_i$ 3-vectors. Considering only the leading 3×3 submatrix of the equation, and multiplying it by its transpose, we get:

$$\mathsf{P}_i\bar{\mathsf{T}}\bar{\mathsf{T}}^\top\mathsf{P}_i^\top \sim \begin{pmatrix} \tau_i & 0 & u_i \\ 0 & 1 & v_i \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} f_i^2 & 0 & 0 \\ 0 & f_i^2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \tau_i & 0 & u_i \\ 0 & 1 & v_i \\ 0 & 0 & 1 \end{pmatrix}^\top,$$

where $\bar{\mathsf{T}}$ is the 4×3 matrix consisting of the first three columns of T. Let

$$\mathsf{X} = \bar{\mathsf{T}}\bar{\mathsf{T}}^\top$$
$$\mathsf{M}_i = \begin{pmatrix} \tau_i & 0 & u_i \\ 0 & 1 & v_i \\ 0 & 0 & 1 \end{pmatrix}^{-1} \mathsf{P}_i.$$

Then, the above equation becomes:

$$\mathsf{M}_i\,\mathsf{X}\,\mathsf{M}_i^\top \sim \begin{pmatrix} f_i^2 & 0 & 0 \\ 0 & f_i^2 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \qquad (1)$$

The matrix X represents the "absolute quadric" (Triggs, 1997), in the space of the projective reconstruction. It is 4×4, symmetric and of rank 3.

Let $\mathbf{m}_{ij}^\top$ be the vector representing the j-th row of M_i. From Eq. (1), the following linear equations on X can be obtained:

$$\begin{aligned} \mathbf{m}_{i1}^\top\,\mathsf{X}\,\mathbf{m}_{i2} &= 0 \\ \mathbf{m}_{i1}^\top\,\mathsf{X}\,\mathbf{m}_{i3} &= 0 \\ \mathbf{m}_{i2}^\top\,\mathsf{X}\,\mathbf{m}_{i3} &= 0 \\ \mathbf{m}_{i1}^\top\,\mathsf{X}\,\mathbf{m}_{i1} - \mathbf{m}_{i2}^\top\mathsf{X}\mathbf{m}_{i2} &= 0. \end{aligned}$$

The rank-3 constraint on X can not be imposed via linear equations, which implies that there exist singularities for the linear method, that are not singular for the generic case (Sturm, 2000). The generic singularities (critical motions) for self-calibration of varying focal length (with other intrinsic parameters known), are described in Kahl et al. (2000) and Sturm (1999). An imaging configuration that is singular for the linear approach, but not in general, is the case where the optical axes of all views pass through one 3D point. Image sequences taken for modeling objects will very often be singular in this respect (e.g. the sequence shown in Fig. 3).

Due to this singularity, the system of the above linear equations will have a one-dimensional family of solutions:

$$\mathsf{X} \sim \mathsf{X}_1 + \mu \mathsf{X}_2.$$

The rank-3-constraint allows to solve for μ via the equation $\det \mathsf{X} = 0$. This is a degree-4-polynomial in μ. We solve it numerically, thus obtaining a maximum of 4 solutions for X. To find a unique solution, we compute the focal lengths that each solution gives rise to, and choose the solution, where these respect practical bounds (they have to lie in an interval of the order [300, 5000], depending on the camera used). In practice, we always found a single solution satisfying these constraints, the others being far off.

Focal lengths are extracted by computing $\omega_i \sim \mathsf{M}_i \mathsf{X} \mathsf{M}_i^\top$ and then

$$f_i = \sqrt{\frac{\frac{1}{2}(\omega_{i,11} + \omega_{i,22})}{\omega_{i,33}}}.$$

From the estimated X, we extract a projective transformation that upgrades projection matrices and point coordinates to metric. There is no unique solution for this, so in practice we choose one that has roughly equal singular values. Let $\mathsf{X} = \pm \mathsf{U} \Sigma \mathsf{U}^\top$ be the singular value decomposition of X. Since X is of rank 3, the 4-th singular value is zero. Let Σ' be obtained by replacing that zero with e.g. the largest singular value in Σ, we obtain the projective upgrade transformation needed:

$$\mathsf{T} = \mathsf{U}\sqrt{\Sigma'}.$$

Extracting extrinsic parameters from the upgraded projection matrices is then straightforward—it basically requires fitting of orthonormal matrices to general 3 × 3 matrices (Horn et al., 1988). More details are given in Section 4.4. The result is optimized via bundle adjustment. An alternative to the described approach would be to use the coplanarity information already available, like Alon and Sclaroff (2000), Malis and Cipolla (2000), Triggs (1998a), Viéville et al. (1995), and Xu et al. (2000).

In the following paragraph, we just give a few numerical details. In order to improve the condition of the linear equation system, we transform the matrices M_i as follows. First, we assume that images are normalized using e.g. Hartley (1995). Second, we make use of the free choice for the basis of the projective reconstruction, by computing a projective transformation, that hopefully leads to better conditioning. A simple method to do that is as follows. We stack the M_i in a matrix M of size $3n \times 4$, and compute its singular value decomposition:

$$\mathsf{M} = \mathsf{A} \Gamma \mathsf{B}^\top.$$

From A, we extract sub-matrices replacing the M_i in the linear equations: A is orthonormal, so the linear equations are more likely to be well conditioned. The product $\Gamma \mathsf{B}^\top$ represents the projective transformation corresponding to the mapping between the original and the transformed M_i (naturally, the 3D points have to be transformed accordingly). Using this normalization, we obtained much more accurate initial values and actually prevented the bundle adjustment to fall in a local minimum it got trapped in otherwise, in one case.

4.2. Structure Parameterization

In this section, we adapt the projective structure parameterization of Section 3.1 to the Euclidean case. In this case, planes are modeled as homogeneous 4-vectors, whereas points can be written as inhomogeneous 3-vectors.

The plane parameterization has been described in Section 3.1.2 and mapped coordinates (cf. Section 3.1.4) were used to reach the minimality. The point parameterization under multi-coplanarity constraints of Section 3.1.3 for the projective case can be used either directly or adapted to take full advantage of the Euclidean structure. We successively specialize the different cases.

Unconstrained Points. As said above, points can be parameterized using inhomogeneous 3-vectors, which is minimal in this case.

Single-Coplanar Points. Let X be a point lying on a plane π. As for the projective case, we want to change the reconstruction basis such as to fix an element of $\mathbf{X}$ to a constant value. In the Euclidean case, we have $\mathbf{X}^\top \sim (\bar{\mathbf{X}}^\top \mid 1)$ in the homogeneous form, so that the 4-th element is already fixed. Consequently, we must choose a transformation that preserves this element while fixing another one. This class of transformation is H_π^j where $j \in \{1 \dots 3\}$. The practical algorithm for parameterizing/unparameterizing such a point in the Euclidean case is similar to that of Table 1 but using the constraint $j \in \{1 \dots 3\}$ for the choice of j.

Multi-coplanar Points. We follow the same reasoning as in the previous case. A point lying on two planes is then parameterized by a scalar and does not have parameters in the three planes case. The practical algorithms are then identical to the projective case, provided a choice for the indices j and j' in $\{1 \dots 3\}$ for the two planes case.

4.3. Motion Parameterization

For motion parameterization in the Euclidean case, we suppose that each camera has z unknown intrinsic parameters, where $z \in \{1 \dots 5\}$.

One View. We choose the reconstruction basis such that $\mathsf{P} \sim \mathsf{K}\ (\mathrm{I} \mid \mathbf{0})$ where K is the calibration matrix, containing the intrinsic parameters. We have therefore z degrees of freedom for this first camera.

Two or More Views. The Euclidean basis has been fixed by the first view up to a global scale factor. We then have to completely parameterize the other camera matrices. Such an additional camera is written as $\mathsf{P}' \sim \mathsf{K}'\ (\mathsf{R} \mid \mathbf{t})$. Making the same assumption on the intrinsic parameters than for the first view, this leaves $z + 6$ degrees of freedom for each view, its internal parameters and the 6 parameters for the rotation R and the translation $\mathbf{t}$. These entities are minimally parameterized, as described in e.g. Atkinson (1996).

4.4. Maximum Likelihood Estimator

The maximum likelihood estimator in the metric case is very similar to that of the projective case as the cost function is the same. The intrinsic parameters for each camera have been recovered previously, see Section 4.1. In order to initialize our parameterization, we still need to extract the relative pose of each camera, i.e. factorize each projection matrix $\mathsf{P} \sim (\bar{\mathsf{P}} \mid \mathbf{p})$ under the form $\mathsf{P} = \frac{1}{\lambda}\mathsf{K}(\mathsf{R} \mid \mathbf{t})$ where λ is an unknown scale factor. Let us define $\mathsf{S} = \mathsf{K}^{-1}\bar{\mathsf{P}}$. We first estimate the scale factor as $\lambda = \sqrt[3]{\det \mathsf{S}}$. The translation can then be obtained by $\mathbf{t} = \lambda\mathsf{K}^{-1}\mathbf{p}$. In the noise-free case, $\lambda\mathsf{S}$ is an orthonormal matrix, but in practice it is not and we choose the closest rotation matrix in the sense of the Frobenius norm. This can be done using a singular value decomposition of $\lambda\mathsf{S}$ and a recomposition where the matrix of singular values Σ is omitted: $\mathsf{R} = \mathsf{U}\mathsf{V}^\top$ where $\lambda\mathsf{S} = \mathsf{U}\Sigma\mathsf{V}^\top$. Once this initialization has been done, non-linear optimization of the cost function $\mathcal{C}$ (cf. Section 3.3) can be launched using the Levenberg-Marquardt algorithm (Gill et al., 1981) with analytic differentiation.

5. Experimental Results Using Simulated Data

In this section, we compare our method to existing ones, notably to that consisting in individually reconstructing each point and to that using approximate multi-coplanarity constraints. We perform this comparison for the structure results, then for the motion results.

The test bench consists of a cube of one meter side length observed by a set of cameras. Points are generated on the cube, possibly offset from their planes in order to simulate non-perfect coplanarity and projected onto the images, where centered Gaussian noise is added. The default parameters of this simulation are the following. Up to 50, 10 and 1 points are generated on respectively each face, edge and vertex of the cube. Two cameras with a focal length of 1000 (expressed in number of pixels) and a 1 meter baseline are situated at a distance of 10 meters from the cube. The standard deviation of image noise is up to 3.0 pixels. The intrinsic parameters are not supposed to be known which yields projective reconstructions.

In the sequel, we vary independently each of these parameters to compare the different approaches under various conditions, especially we want to determine how the constrained methods behave when the observed surfaces are only approximately planar.

We measure the quality of reconstructions using the 3D residual of its Euclidean distance to the ground truth scene structure $\underline{X}$: $E = \sqrt{\frac{1}{m}\sum_{j=1}^{m} d^2(\mathsf{H}\mathbf{X}_j, \underline{\mathbf{X}}_j)}$, where H is a 3D homography (mapping the projective to the Euclidean structure) estimated using non-linear minimization of E. We measure the median value over 100 trials.

The estimators compared are:

- *Po-ML*: Optimal structure and motion obtained in a bundle adjustment manner (Triggs et al., 2000) without geometric constraints;
- *Pl-wt*: (*wt* stands for weights) similar to *Po-ML* but uses heavily weighted ($2^{60} \approx 10^{20}$) additional equations to approximate multi-coplanarity (McGlone, 1996; Szeliski and Torr, 1998);
- *Pl-ML*: Uses the parameterization described in this paper to explicitly model multi-coplanarity;
- *Pl-h*: (*h* stands for homography) uses method *Po-ML* described above with as input point correspondences corrected by maximum likelihood estimation of homographies. This method is described in more detail below. Note that it works only with two images and with the single-coplanarity constraint.

The last method evaluated relies on a simple homography-based point correction. A plane observed by two cameras induces an homography. This homography relates the projections of the points lying on the plane. The family of such homographies is 3-dimensional, provided that the epipolar geometry is known (this is linked to the fact that a plane has 3 degrees of freedom). In the calibrated case, they depend upon the relative pose between the two cameras and on their intrinsic parameters. If all these consistency constraints are ignored, and if the piecewise planar structure and motion problem is considered only for two views and with single-coplanarity constraints, one can devise a simple process to incorporate the knownledge of coplanarity in a standard unconstrained reconstruction method. Indeed, one can estimate independently each homography corresponding to each coplanar group of points and correct them so that they perfectly correspond through the homography. A standard structure and motion algorithm can then be launched with as input the corrected points. This is what *Pl-h* does. Obviously, this process is suboptimal since most consistency constraints have been ignored, and since the final reconstruction is only approximately planar. Extending the idea to multi-view and multi-coplanarity constraints, by enforcing all the underlying consistency constraints would yield the same result as our estimator, up to the convergence of the underlying non-linear optimizers. However, the algebraic structure would be more complicated since more consistency constraints have to be imposed in the images than in the 3D space.

Let us describe the different experimental situations when varying a scene parameter and the simulation results we have obtained.

Added Image Noise (Fig. 1(a)): The standard deviation of added image noise is varied from 0 to 3 pixels;

Baseline (Fig. 1(b)): The baseline is varied between 0.1 and 1 meter;

Number of Points (Fig. 1(c)): The number of points is respectively equal to 50α, 10α and 1 for each face, edge and vertex of the cube, where α varies from 0.1 to 1;

Number of Views (Fig. 1(d)): The number of views varies from 2 to 10. The different cameras are situated such that the baseline between two consecutive ones is 1 meter;

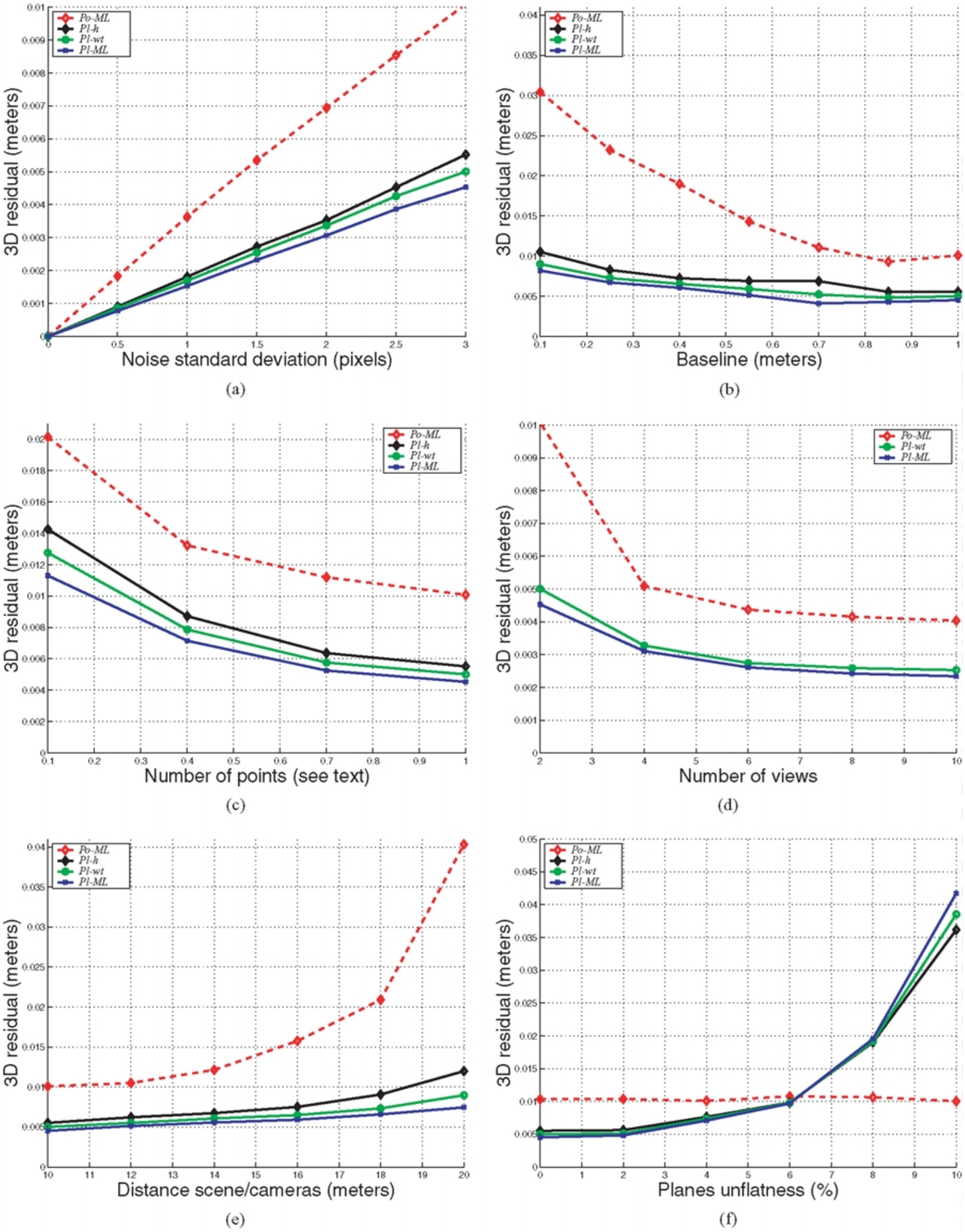
Distance Scene/Cameras (Fig. 1(e)): The distance between the cube and the cameras is varied between 10 and 20 meters.

In all these cases, the method *Po-ML* based only on individual point reconstruction gives results of a quality lower than methods *Pl*-modeling also planes (the residual is at least twice as low). The method *Pl-ML* performs slightly better than *Pl-wt* in all cases. Finally, method *Pl-h* gives results slightly worse than *Pl-wt*, but much better than *Po-ML*.

One aspect not shown on the graphs of Fig. 1, due to the use of a median value over a large number of trials, is that methods *Po-ML* and *Pl-wt* have a percentage of convergence lower than *Pl-ML* and *Pl-h*, especially for unstable configurations (large image noise, small baseline, high distance scene/cameras etc.). For example, the percentage of convergent estimations over 1000 trials is 95.2%, 89.1%, 97.5% and 97.3% for *Po-ML*, *Pl-wt*, *Pl-ML* and *Pl-h* respectively, for a distance scene/cameras of 20 meters and a 0.1 meter baseline.

Plane Unflatness (Fig. 1(f)): 3D points are offset from the planes they lie on by distances drawn from a normal distribution with standard deviation between 0 and 0.1 meters.

We observe that there is a threshold on the plane unflatness where methods *Pl-* using the knowledge of planes begin to perform worse than method *Po-ML*. It is interesting to define the *breakdown ratio*, denoted by ε, as the ratio between the extent of 3D noise and plane surface area, assuming that the scene is seen completely in all views. In the case of Fig. 1(f),



(a)

(b)

(c)

(d)

(e)

(f)

Figure 1. Comparison of the 3D residuals for different approaches versus different scene parameters. Note that method *Pl-h* is not visible on (d) since it works with two views only.

Table 5. Breakdown ratio ε for different scene configurations (image noise, number of views, distance scene/cameras).

	3 m (%)	10 m (%)	20 m (%)
$n = 2$			
1 pixel	0.5	2	4
3 pixels	2	6	9
$n = 10$			
1 pixel	0.3	1.2	3
3 pixels	1.3	4	8

$\varepsilon = 6\%$, recalling that each plane of the cube is 1 square meter. The value of ε depends on all scene parameters.

Table 5 shows values of ε established experimentally for various scene parameters. We observe that the less stable the configuration is the higher is ε, i.e. the more important is the incorporation of multi-coplanarity constraints, even if the scene is not perfectly piecewise planar.

The values of one or several percent in Table 5 represent relatively large variations which are superior to those of a great majority of approximately planar real surfaces. Consequently, we can say that there are many cases when a method using piecewise planarity will perform better than any method based on individual point reconstruction.

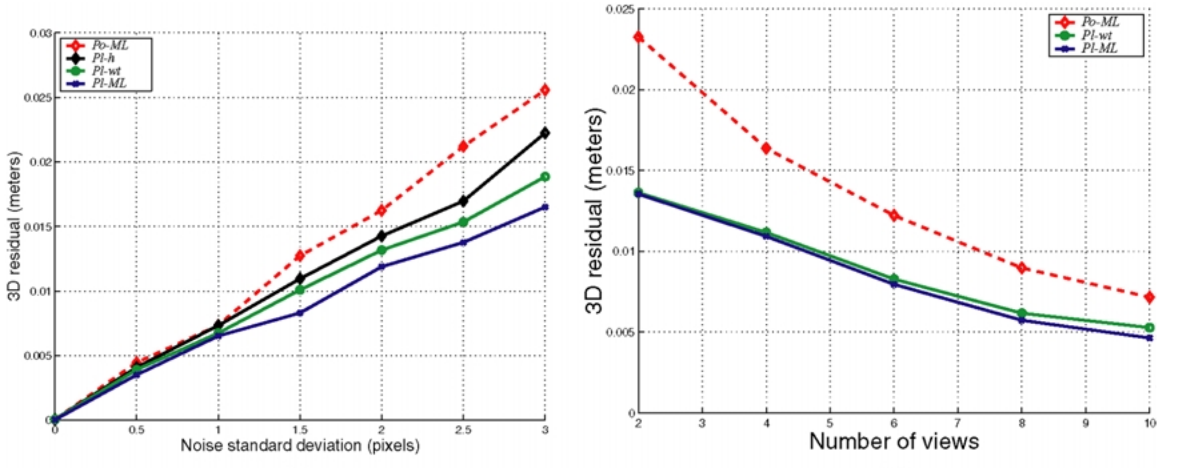
Similar results with other point- and plane-based methods have been obtained in Bartoli and Sturm (2001). We have also performed similar experiments in the calibrated case, i.e. the reconstructions obtained are Euclidean, and we observed that this does not change the results significantly. This can be explained by the fact that the optimization criterion is image-based, and so invariant to projective transformation (such as the upgrade from projective to metric space).

Comparing the Motion Estimates. We compare the results on the motion parameters provided by the different methods. We use the same experimental setup as previously. The quality of the estimated motion is measured as follow. We extract the n projection centers C_i of the estimated camera matrices and compute the 3D residual of their Euclidean distances to the ground-truth projection centers $\underline{C}_i$: $E_{\text{motion}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d^2(\mathsf{H}C_i, \underline{C}_i)}$. The 3D homography H is estimated as in the previous case, using non-linear minimization of E, i.e. using estimated to ground-truth point correspondences (estimating it with corresponding centers of projection would be highly sensitive to noise, due to the low number of data). We measure the median value of E_{motion} over 100 trials.

Let us describe the different experimental situations and results obtained.

Added Image Noise (Fig. 2(a)): The standard deviation of added image noise is varied from 0 to 3 pixels;

Number of Views (Fig. 2(b)): The number of views is varied from 2 to 10, a 3 pixels standard deviation noise is added.



(a)

(b)

Figure 2. Comparison of the 3D residuals for the motion for different approaches versus different scene parameters. Note that method *Pl-h* is not visible on (b) since it works with two views only.

As already observed for the results on the structure, the method *Po-ML* that do not use coplanarity information performs worse than the others. The method *Pl-ML* performs better than *Pl-wt* and the method *Pl-h* performs worse than *Pl-wt*. We observe that the gap between plane-based methods *Pl-* and the point-based method *Po-ML* is reduced compared to the error estimated on the structure. In all cases, we also observe that the error measure obtained is worse than for the structure. This is due to the fact that the homography mapping the reconstruction result to the ground-truth data is estimated by minimizing the criterion E, based on the structure only.

6. Results Using Real Images

In this section, we present the reconstruction results obtained using the images shown in Fig. 3. Similar results have been obtained with other images (see Bartoli et al., 2001). We describe the different steps followed to perform a complete reconstruction, from the images to the 3D textured model. Table 6 shows the reprojection errors obtained at various stages of the process.

Structure and Motion Initialization. This has been obtained using image point matches given manually. We perform a partial reconstruction from two images using the method (Hartley, 1995; Hartley and Sturm, 1997) and incrementally add the other images to obtain the complete structure and motion. We then run a bundle adjustment to minimize the reprojection error and to obtain the maximum likelihood estimate for an unconstrained structure.

Multi-coplanarity Constraints. These relationships are established semi-automatically using plane homographies. The user provides three image points matched in at least one other view to obtain a first guess for the plane. The other points lying on this plane are then automatically detected. The user may interact to correct badly clustered points and add points visible in only one view.

Table 6. Reprojection errors (pixel) and number of iterations of non-linear optimizers at various stages of the reconstruction process. MLE stands for Maximum Likelihood Estimator.

Space	Approach	Step	Rep. error (pixels)	No. of iterations
Projective	Unconstrained	Init.	3.86	–
		MLE	1.07	7
	Constrained	Init.	1.90	–
		MLE	1.20	3
Metric	Unconstrained	MLE	2.69	6
	Constrained	Init.	3.86	–
		MLE	1.43	9

Constrained Refinement of Structure and Motion. From the previous data, the structure is parameterized as described in Section 3 and the maximum likelihood estimate for constrained structure and motion of Section 3.3 is computed. According to Table 4, points visible in only one view and constrained to lie on two or more planes are reconstructed and involved in the optimization process.

Structure Completion. Points appearing in only one view and lying on one plane are then reconstructed. The structure is complete in the sense that no more points will be further added. Figure 5 shows structure reprojection on an original image.

Calibration. The metric structure is obtained via the self-calibration process described in Section 4.1 and the reprojection error is minimized while enforcing the multi-coplanarity constraints as indicated in Section 4.4. Figure 4 shows different views of the recovered structure and the positioning of the cameras and Fig. 5 the reprojection of the model in two original images. For the intrinsic parameters of each camera, only the focal lengths are involved. It appears that also including principal points does not change significantly the results.

Texture Maps. The texture mapping requires the user to provide a polygonal delineation for each planar facet

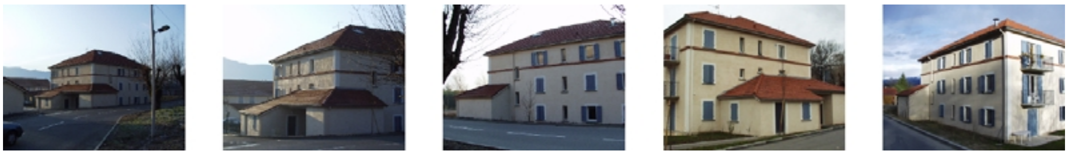


Figure 3. Images used to validate the method.

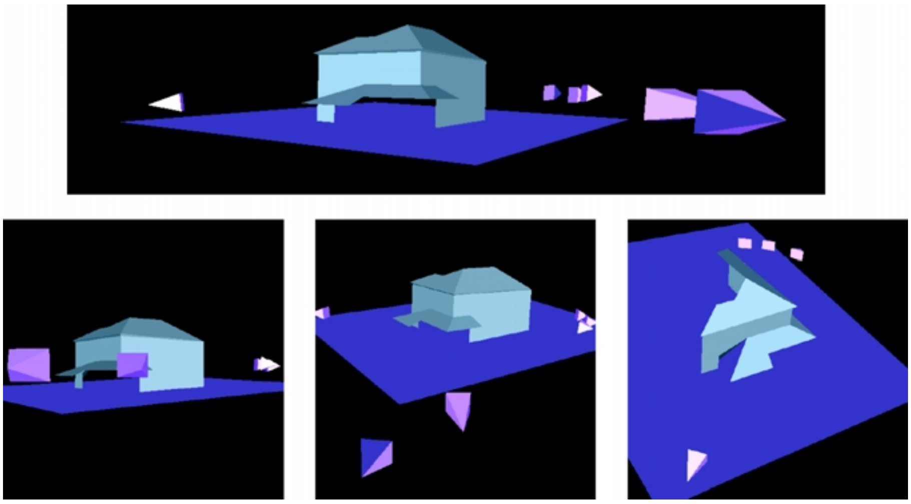


Figure 4. Recovered metric structure and motion. The structure is shown as a set of planar polygons while the different cameras (the motion) are represented by pyramids. The height of a pyramid is proportional to the recovered focal length of the camera. The bottom-right image shows a rendering from above the point of view of the right image of Fig. 5.

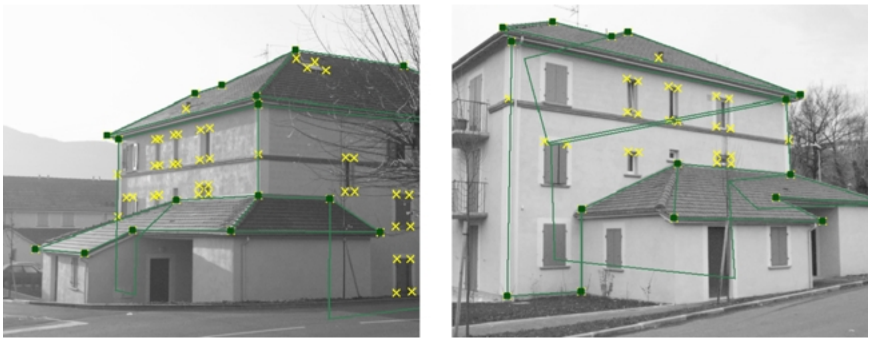


Figure 5. Reprojection of the recovered model onto the original images. The yellow crosses indicate the position of the reprojected point features.

in one of the images. The texture maps are then extracted and perspectively corrected using calibrated projection matrices and bicubic interpolation. Figure 6 shows different views of the recovered textured model.

Quality Assessment. We have performed several measures on the metric reconstruction "before" and "after" the constrained optimization process (i.e. reflecting the changes when using the method described in this paper). Two kinds of quantity are significant: length ratios and angles. Table 7 shows measures of such quantities. In this table, σ_1 and σ_2 are the variances of the length of respectively the height and width of the largest windows on the two walls, whereas μ is the mean of $1 - 2\alpha_i/\pi$ where α_i are the measures of right angles. The values given in Table 7 show that the metric reconstruction obtained with our method is clearly of superior quality than the unconstrained one.

Table 7. Metric measures on the Euclidean reconstruction "before" and "after" the constrained optimization. The lower λ_1, λ_2 and μ (see text) are, the better the reconstruction is.

	σ_1	σ_2	μ
Before	0.0489	0.0254	0.1032
After	0.0102	0.0168	0.0720

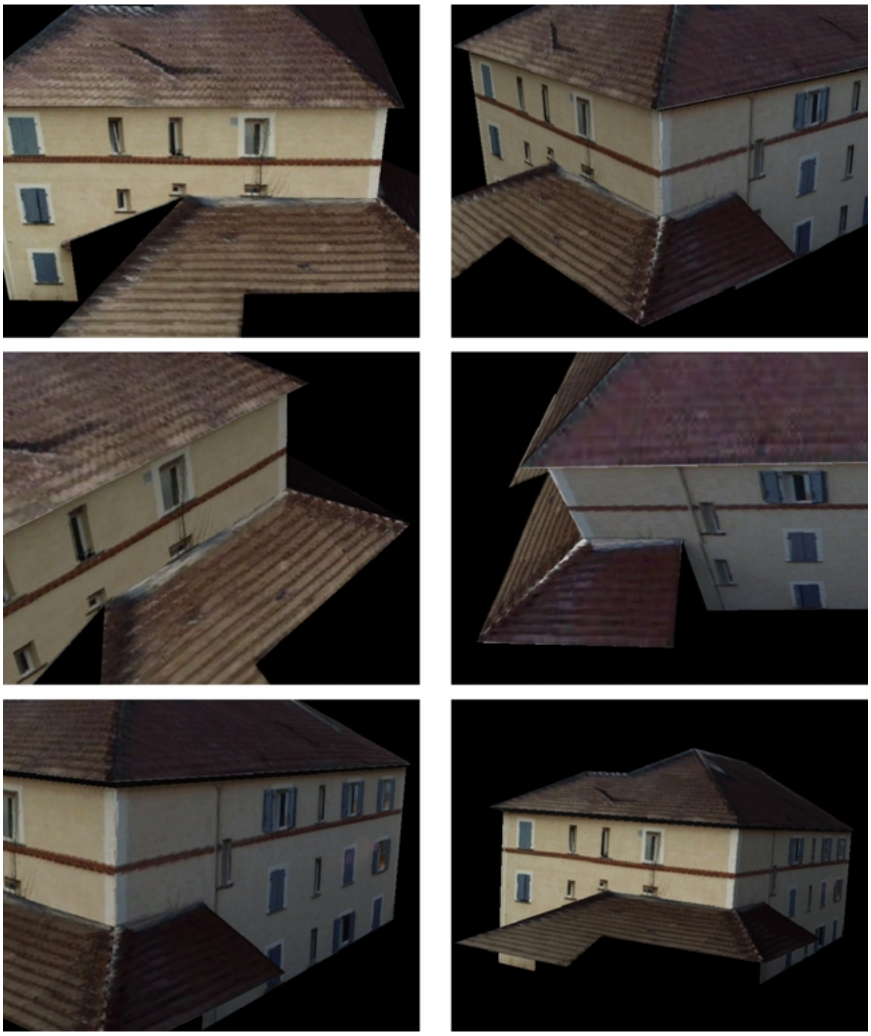


Figure 6. Different views of the textured model. Note that artefacts may be induced by possibly unmodeled non-planar parts of the surfaces, e.g. the pole bulging out of the roof in the top-right image is wrongly mapped to the roof plane, and is therefore distorted in other views, e.g. the top-left one.

7. Conclusions

We have presented an hybrid approach that draws on the strengths of both the traditional feature- and primitive-based approaches, i.e. the reconstruction is accurate and the recovered model allows to produce photorealistic rendering. More precisely, we focus on the case of points and planes related by multi-coplanarity constraints and on the design of a *constrained structure and motion* maximum likelihood estimator in both the projective and the metric cases. This maximum likelihood estimator uses a *minimal* parameterization of scene structure, enforcing underlying geometric constraints and a quasi-minimal parameterization of motion.

Experimental results on simulated data show that the quality of the reconstruction obtained with our method is clearly superior to those of traditional feature-based methods, in a large variety of experimental configurations, and for both structure and motion. We also consider the case when surfaces are only approximately planar and experimentally determined breakpoints of plane unflatness above which the incorporation of multi-coplanarity constraints makes the estimation less reliable.

The method is validated using real images. The results are convincing, in terms of both rendering quality and accuracy of metric values compared to a feature-based method.

The implementation of our methods comprises modules for unconstrained projective reconstruction ("linear" ones and bundle adjustment), constrained projective reconstruction (initialization and optimization), self-calibration ("linear" method and optimization), as well as constrained Euclidean reconstruction (initialization and bundle adjustment).

Notes

1. Note that this is very different from the hybrid approach of Debevec et al. (1996) which is actually primitive-based.
2. This is not true if the planes form a pencil.

References

Alon, J. and Sclaroff, S. 2000. Recursive estimation of motion and planar structure. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, South Carolina, USA, pp. 550–556.

Atkinson, K.B. (Ed.). 1996. *Close Range Photogrammetry and Machine Vision*. Whittles Publishing.

Baillard, C. and Zisserman, A. 1999. Automatic reconstruction of piecewise planar models from multiple views. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Fort Collins, Colorado, USA, IEEE Computer Society Press: Los Alamitos, CA, pp. 559–565.

Bartoli, A. and Sturm, P. 2001. Constrained structure and motion from *N* views of a piecewise planar scene. In *Proceedings of the First International Symposium on Virtual and Augmented Architecture (VAA'01)*, Dublin, Ireland, pp. 195–206.

Bartoli, A., Sturm, P., and Horaud, R. 2001. Projective structure and motion from two views of a piecewise planar scene. In *Proceedings of the 8th International Conference on Computer Vision*, Vancouver, Canada, Vol. 1, pp. 593–598.

Beardsley, P., Torr, P., and Zisserman, A. 1996. 3D model acquisition from extended image sequences. In *Proceedings of the 4th European Conference on Computer Vision*, Cambridge, England, B. Buxton and R. Cipolla (Eds.), Vol. 1065 of Lecture Notes in Computer Science. Springer-Verlag: Berlin, pp. 683–695.

Berthilsson, R. and Heyden, A. 1998. Recognition of planar point configurations using the density of affine shape. In *Proceedings of the 6th International Conference on Computer Vision*, Bombay, India, pp. 72–88.

Cross, G. and Zisserman, A. 2000. Surface reconstruction from multiple views using apparent contours and surface texture. In *NATO Advanced Research Workshop on Confluence of Computer Vision and Computer Graphics*, Ljubljana, Slovenia, A. Leonardis, F. Solina, and R. Bajcsy (Eds.), pp. 25–47.

Debevec, P.E., Taylor, C.J., and Malik, J. 1996. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *SIGGRAPH'96*, New Orleans.

Dick, A.R., Torr, P.H.S., Ruffle, S.F., and Cipolla, R. 2001. Combining single view recognition and multiple view stereo for architectural scenes. In *Proceedings of the 8th International Conference on Computer Vision*, Vancouver, Canada.

Faugeras, O. and Lustman, F. 1988. Motion and structure from motion in a piecewise planar environment. *International Journal of Pattern Recognition and Artificial Intelligence*, 2(3):485–508.

Fornland, P. and Schnörr, C. 1997. A robust and convergent iterative approach for determining the dominant plane from two views without correspondence and calibration. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Puerto Rico, USA, IEEE (Ed.), pp. 508–513.

Gill, P.E., Murray, W., and Wright, M.H. 1981. *Practical Optimization*. Academic Press: New York.

Gortler, S.J., Grzeszczuk, R., Szeliski, R., and Cohen, M. 1996. The lumigraph. In *Proceedings of SIGGRAPH*, New Orleans, LA, pp. 43–54.

Hartley, R. 1995. In defence of the 8-point algorithm. In *Proceedings of the 5th International Conference on Computer Vision*, Cambridge, MA, USA, pp. 1064–1070.

Hartley, R. and Sturm, P. 1997. Triangulation. *Computer Vision and Image Understanding*, 68(2):146–157.

Heyden, A. and Åström, K. 1995. A canonical framework for sequences of images. In *Workshop on Representation of Visual Scenes*, Cambridge, MA, USA, pp. 45–52.

Heyden, A. and Åström, K. 1998. Minimal conditions on intrinsic parameters for euclidean reconstruction. In *Proceedings of the Third Asian Conference on Computer Vision*, Hong Kong, Vol. II, pp. 169–176.

Horn, B.K.P., Hilden, H.M., and Negahdaripour, S. 1988. Closed-form solution of absolute orientation using orthonormal matrices. *Journal of the Optical Society of America A*, 5(7):1127–1135.

Kahl, F., Triggs, B., and Åström, K. 2000. Critical motions for auto-calibration when some intrinsic parameters can vary. *Journal of Mathematical Imaging and Vision*, 13(2):131–146.

Kutulakos, K.N. and Seitz, S.M. 1999. A theory of shape by space carving. In *Proceedings of the 7th International Conference on Computer Vision*, Kerkyra, Greece, Vol. 1, pp. 307–314.

Lang, F. and Förstner, W. 1996. 3D-city modeling with a digital one-eye stereo system. In *Proceedings of the XVIII ISPRS-Congress*, Vienna, Austria.

Levoy, M. and Hanrahan, P. 1996. Light field rendering. In *Proceedings of SIGGRAPH*, New Orleans, LA, pp. 31–42.

Malis, E. and Cipolla, R. 2000. Multi-view constraints between collineations: Application to self-calibration from unknown planar structures. In *Proceedings of the 6th European Conference on Computer Vision*, Dublin, Ireland, D. Vernon (Ed.), Vol. 1843 of Lecture Notes in Computer Science. Springer-Verlag: Berlin, pp. 610–624.

Maybank, S.J. and Faugeras, O.D. 1992. A theory of self calibration of a moving camera. *International Journal of Computer Vision*, 8(2):123–151.

McGlone, C. 1996. Bundle adjustment with geometric constraints for hypothesis evaluation. In *Proceedings of the XVIII ISPRS-Congress*, Vienna, Austria, pp. 529–534.

Niem, W. 1994. Robust and fast modelling of 3D natural objects from multiple views. In *Proceedings of the SPIE Conference on*

Image and Video Processing II, San Jose, USA, Vol. 2182, pp. 388–397.

Pollefeys, M. 1999. Self-calibration and metric 3D reconstruction from uncalibrated image sequences. Ph.D. Thesis, Katholieke Universiteit Leuven, Belgium, Faculteit Toegepaste Wetenschappen, Arenbergkasteel, B-3001 Heverlee, Belgium.

Pollefeys, M., Koch, R., and Van Gool, L. 1998. Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. In *Proceedings of the 6th International Conference on Computer Vision*, Bombay, India, pp. 90–95.

Pollefeys, M., Vergauwen, M., and Van Gool, L. 2000. Automatic 3D modeling from image sequences. In *Proceedings of the XIX ISPRS-Congress*, Amsterdam, the Netherlands, Vol. B5, pp. 619–626.

Seitz, S.M. and Dyer, C.R. 1997. Photorealistic scene reconstruction by voxel coloring. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Puerto Rico, USA, IEEE Computer Society Press: Los Alamitos, CA, pp. 1067–1073.

Sinclair, D. and Blake, A. 1996. Quantitative planar region detection. *International Journal of Computer Vision*, 18(1):77–91.

Slama, C.C. (Ed.). 1980. *Manual of Photogrammetry*, Fourth edn. American Society of Photogrammetry and Remote Sensing: Falls Church, Virginia, USA.

Streilein, A. and Hirschberg, U. 1995. Integration of digital photogrammetry and CAAD: Constraint-based modeling and semi-automatic measurement. In *Proceedings of the International CAAD Futures Conference*, Singapore.

Sturm, P. 1999. Critical motion sequences for the self-calibration of cameras and stereo systems with variable focal length. In *Proceedings of the Tenth British Machine Vision Conference*, Nottingham, England, T. Pridmore and D. Elliman (Eds.), British Machine Vision Association, pp. 63–72.

Sturm, P. 2000. A case against Kruppas equations for camera self-calibration. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1199–1204.

Sturm, P. and Triggs, B. 1996. A factorization based algorithm for multi-image projective structure and motion. In *Proceedings of the 4th European Conference on Computer Vision*, Cambridge, England, B. Buxton and R. Cipolla (Eds.), Vol. 1065 of Lecture Notes in Computer Science. Springer-Verlag: Berlin, pp. 709–720.

Szeliski, R. 1993. Rapid octree construction from image sequences. *Computer Vision, Graphics and Image Processing*, 58(1):23–32.

Szeliski, R. and Torr, P.H.S. 1998. Geometrically constrained structure from motion: Points on planes. In *3D Structure from Multiple Images of Large-Scale Environments (SMILE'98)*. Springer Verlag: Berlin.

Tarel, J.-P. and Vézien, J.-M. 1995. A generic approach for planar patches stereo reconstruction. In *Proceedings of the Scandinavian Conference on Image Analysis*, Uppsala, Sweden, pp. 1061–1070.

Triggs, B. 1997. Autocalibration and the absolute quadric. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Puerto Rico, USA, IEEE Computer Society Press: Los Alamitos, CA, pp. 609–614.

Triggs, B. 1998a. Autocalibration from planar scenes. In *Proceedings of the 5th European Conference on Computer Vision*, Freiburg, Germany.

Triggs, B. 1998b. Optimal estimation of matching constraints. In *3D Structure from Multiple Images of Large-Scale Environments (SMILE'98)*, Springer Verlag: Berlin.

Triggs, B., McLauchlan, P.F., Hartley, R.I., and Fitzgibbon, A. 2000. Bundle ajustment—A modern synthesis. In *Proceedings of the International Workshop on Vision Algorithms: Theory and Practice*, Corfu, Greece, B. Triggs, A. Zisserman, and R. Szeliski (Eds.), Vol. 1883 of Lecture Notes in Computer Science. Springer-Verlag: Berlin, pp. 298–372.

Viéville, T., Zeller, C., and Robert, L. 1995. Using collineations to compute motion and structure in an uncalibrated image sequence. *International Journal of Computer Vision*, 20(3):213–242.

Xu, G., Terai, J.-I., and Shum, H.-Y. 2000. A linear algorithm for camera self-calibration, motion and structure recovery for multi-planar scenes from two perspective images. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, South Carolina, USA.

Zhang, Z. 1998. Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision*, 27(2):161–195.