Projective Translations and Affine Stereo Calibration

Andreas Ruf Gabriela Csurka Radu Horaud

GRAVIR-IMAG, INRIA Rhone-Alpes
655, avenue de l'Europe
38330 Montbonnot St.Martin, France

Abstract

This paper investigates the homography which transforms a set of points in projective space when undergoing a rigid translation, termed a *projective translation*. A representation with seven parameters is proposed. They represent explicitly the geometric entities constraining and defining the translation. A practical algebraic method for extracting these parameters is developed. It allows for affine calibration of a stereo rig, for characterizing the motion intrinsics, as well as for composing projective translations. The practical effectiveness of calibration is evaluated on synthetic and real image data.

1 Introduction

The recovery of structure and motion is one of the essential problems in machine vision. Difficulties increase when uncalibrated cameras are considered, since in this case metric information is missing. Nevertheless, rigidity of motions observed with an uncalibrated stereo camera imposes strong constraints on the transforms describing them. They allow for an augmented representation of structure and motion. More precisely, the homography which maps a sets of points reconstructed in projective space from their initial to their final position is algebraically similar or "conjugate" to this displacement [11], [2]. Representation of structure up to scaled-Euclidean is possible.

The similarity class of projective translations is analyzed and an intrinsic parameterization is introduced. In practice, estimates of these matrices always deviate from the theoretical form due to round-off errors, measurement noise, and outliers caused by mismatches. To overcome this, a computational method is proposed that decomposes into the proposed form, that is numerically stable, that robustly cumulates several inputs, and that has a number of important practical applications.

Replacing classical calibrated systems by their uncalibrated successors is an important issue, since such systems exhibit higher efficiency, autonomy, and flexibility. These systems demand just the level of calibration necessary for a task and perform self-calibration with minimum or without a-priori knowledge. In our case, self-calibration amounts to affine stereo calibration from unknown translations.

Translations are an important class of motions, because they frequently occur in practice, they are easy to implement, and they highly facilitate correlation based matching and tracking. We demonstrate that even the "uncalibrated" representation allows to extract the intrinsic properties of translation: the plane at infinity which ensures parallel traces, the direction or axis, and covered parallel distance.

Practical applications exploiting this knowledge are numerous and relevant. Firstly, the plane at infinity allows to upgrade the rather poor projective representation of structure to its affine representation. Secondly, the direction allows inter-/extrapolating and superposing translational motions. Finally, the distance allows to do so uniformly or to localize on the axis. Potential real-world applications worth noting in this especial context are vision-based navigation of autonomous robots, visual servoing of robot manipulators, and a-priori or reactive task planning.

Previous work on affine calibration of a stereo image pair used constructive methods based on projective invariants [10], [9] or the detection of three image points arising from points lying in the plane at infinity. These points can be either three vanishing points associated with translational motions [4], [9] or three virtual points associated with ground-plane motions [1]. The advantage of our method with respect to these methods is that it considers any number of image points, possibly a large one, arising from 3-D points lying in general position – they must not necessarily lie on the plane at infinity. Therefore our method does not rely on special points which are, sometimes, difficult to observe. The fact that a large number of points can be used (and not just three points) inforces the numerical stability of the solution.

This work contributes to the field of uncalibrated vision by considering the projective representation of translational motion, by parameterizing it with the intrinsic variables, and by developing a stable and robust method for extracting these. Numerous applications are sketched and the effectiveness and practicability of affine stereo calibration is experimentally evaluated on synthetic and real image data.

1.1 Paper organization

Some fundamentals and scope are given in section 2. The 3. section rigorously characterizes the algebraic structure of a projective translation. The 4. section proposes a new parameterization and geometrically interprets it. In section 5, a computational method for cumulative decomposition is developed and applications are sketched. Section 6 experimentally evaluates affine stereo-calibration.

1.2 Notations

Matrices and vectors are typeset in boldface. Vectors are columns and row vectors are written by transposing a column. Plain types in Roman designate coordinates, Greek letters designate scale factors. Among 4-by-4 matrices, homographies of $\mathcal{E}$ Euclidean, $\mathcal{A}$ affine, and $\mathcal{P}$ projective three-space are written as $\boldsymbol{T}$, $\boldsymbol{A}$, and $\boldsymbol{H}$. Points in these spaces are represented by vectors $\boldsymbol{X} = [X, Y, Z, 1]^T$, $\boldsymbol{N} = [U, V, W, 1]^T$, $\boldsymbol{M} = [U, V, W, T]^T$, and ideal points by $\boldsymbol{X}_\infty = [X, Y, Z, 0]^T$ or $\boldsymbol{N}_\infty = [U, V, W, 0]^T$. Image points are in pixel coordinates $\boldsymbol{m} = [u, v, 1]^T$. All coordinates are thought as homogeneous.

2 Fundamentals

A camera of the pinhole model [3] projects scene points $\boldsymbol{X}$ in the canonical *Euclidean camera frame* $\mathcal{E}$ onto image pixels $\boldsymbol{m} = \boldsymbol{P}\boldsymbol{X}, \boldsymbol{P} = \begin{bmatrix}\boldsymbol{K}\boldsymbol{0}\end{bmatrix}$, where $\boldsymbol{K}$ holds parameters of the *intrinsic geometry*.

A canonical *affine camera frame* $\mathcal{A}$ is defined by $\boldsymbol{K}$, such that points $\boldsymbol{N} = [U, V, W, 1]^T$ in $\mathcal{A}$ are projected trivially by $\boldsymbol{P}_I = \begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\end{bmatrix}$

$$\boldsymbol{m} = \boldsymbol{P}_I\, \boldsymbol{N}, \quad \boldsymbol{N} = \begin{bmatrix}\boldsymbol{K} & \boldsymbol{0}\\ \boldsymbol{0}^T & 1\end{bmatrix}\boldsymbol{X} \tag{1}$$

The inversion is called *affine-Euclidean upgrade*

$$\boldsymbol{X} = \boldsymbol{A}_{\mathcal{E}A}\boldsymbol{N}, \quad \boldsymbol{A}_{\mathcal{E}A} = \begin{bmatrix}\boldsymbol{K}^{-1} & \boldsymbol{0}\\ \boldsymbol{0}^T & 1\end{bmatrix}. \tag{2}$$

The *extrinsic stereo geometry* is given by the transform $\boldsymbol{T}$ from right to left camera frame, $\mathcal{E}$ to $\mathcal{E}'$

$$\boldsymbol{X}' = \boldsymbol{T}\boldsymbol{X}, \quad \boldsymbol{T} = \begin{bmatrix}\boldsymbol{R} & \boldsymbol{t}\\ \boldsymbol{0}^T & 1\end{bmatrix}. \tag{3}$$

In connection with the left intrinsic parameters $\boldsymbol{K}'$, the *affine stereo geometry* is the transform $\boldsymbol{A}$ from right to left affine camera frame, $\mathcal{A}$ to $\mathcal{A}'$,

$$\boldsymbol{N}' = \boldsymbol{A}\boldsymbol{N}, \quad \boldsymbol{A} = \begin{bmatrix}\boldsymbol{K}'\boldsymbol{R}\boldsymbol{K}^{-1} & \boldsymbol{K}'\boldsymbol{t}\\ \boldsymbol{0} & 1\end{bmatrix}, \tag{4}$$

where $\boldsymbol{e}' = \boldsymbol{K}'\boldsymbol{t}$ is the left epipole and

$$\boldsymbol{H}_\infty = \boldsymbol{K}'\boldsymbol{R}\boldsymbol{K}^{-1} \tag{5}$$

is the *infinity homography* that maps points at infinity from right to left.

The corresponding projections into the left camera are

$$\boldsymbol{P}_{\mathcal{E}} = \begin{bmatrix}\boldsymbol{K}'\boldsymbol{R} & \boldsymbol{K}'\boldsymbol{t}\end{bmatrix}, \quad \boldsymbol{P}_{\mathcal{A}} = \begin{bmatrix}\boldsymbol{H}_\infty & \boldsymbol{e}'\end{bmatrix} \tag{6}$$

The *epipolar stereo geometry* is represented by the fundamental matrix $\boldsymbol{F}$. From $\boldsymbol{F}$ alone, a projection matrix is calculated [5]

$$\boldsymbol{P}_{\mathcal{P}} = \begin{bmatrix}\boldsymbol{H}_\infty & \boldsymbol{0}\end{bmatrix} + \boldsymbol{e}'\boldsymbol{a}^T, \tag{7}$$

where $\boldsymbol{a}^T$ is a row vector of 4 arbitrary elements.

Depending on which matrix $\boldsymbol{P}$ is used for triangulation [6], the obtained *reconstructions* are respectivly in the frames $\mathcal{E}$, $\mathcal{A}$, or $\mathcal{P}$, and are called Euclidean, affine, or projective [4].

The *projective-affine upgrade* from $\mathcal{P}$ to $\mathcal{A}$ is by

$$\boldsymbol{N} = \boldsymbol{H}_{\mathcal{A}P}\boldsymbol{M}, \quad \boldsymbol{H}_{\mathcal{A}P} = \begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \multicolumn{2}{c}{\boldsymbol{a}^T}\end{bmatrix}, \tag{8}$$

since $\boldsymbol{P}_I = \boldsymbol{P}_I\boldsymbol{H}_{\mathcal{A}P}^{-1}, \quad \boldsymbol{P}_{\mathcal{A}} = \boldsymbol{P}_{\mathcal{P}}\boldsymbol{H}_{\mathcal{A}P}^{-1}$.

Points $\boldsymbol{M}_\infty$ in $\mathcal{P}$ satisfying

$$\boldsymbol{N}_\infty = \boldsymbol{H}_{\mathcal{A}P}\boldsymbol{M}_\infty, \text{i.e. } \ \boldsymbol{a}^T\ \boldsymbol{M}_\infty = 0 \tag{9}$$

map onto points $\boldsymbol{N}_\infty$ at infinity $\boldsymbol{N}_\infty$, so $\boldsymbol{a}^T$ is the equation of the *hyper-plane at infinity* $\boldsymbol{\pi}_\infty^T$ in the *projective camera frame* $\mathcal{P}$. The one-step *Projective-Euclidean upgrade* is

$$\boldsymbol{X} = \boldsymbol{H}_{\mathcal{E}P}\boldsymbol{M}, \quad \boldsymbol{H}_{\mathcal{E}P} = \begin{bmatrix}\boldsymbol{K}^{-1} & \boldsymbol{0}\\ \multicolumn{2}{c}{\boldsymbol{a}^T}\end{bmatrix}. \tag{10}$$

Points at infinity $\boldsymbol{X}_\infty = [X, Y, Z, 0]^T$ represent directions. They are the *vanishing points* of translations by $\boldsymbol{t} = [X, Y, Z]^T$. Henceforth, we suppose that a weakly calibrated stereo system with constant intrinsic and extrinsic geometry is reconstructing the scene structure at time instants $i \in (0, 1, 2, \ldots)$ in the projective frame $\mathcal{P}$. Suppose also that inbetween i and $i+1$ the corresponding points $\boldsymbol{X}_i$ or $\boldsymbol{M}_i$ undergo an arbitrary rigid motion

$\boldsymbol{X}^{i+1} = \boldsymbol{T}_{RT}^{(i)}\boldsymbol{X}^{(i)}$. Then, there exists a homography $\boldsymbol{H}_{RT}^{(i)}$ of $\mathcal{P}$, such that

$$\boldsymbol{H}_{RT}^{(i)} = \gamma \begin{bmatrix} \boldsymbol{K}^{-1} & \boldsymbol{0} \\ \multicolumn{2}{c}{\boldsymbol{a}^T} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{R}^{(i)} & \boldsymbol{t}^{(i)} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{K}^{-1} & \boldsymbol{0} \\ \multicolumn{2}{c}{\boldsymbol{a}^T} \end{bmatrix}, \tag{11}$$

Since $\boldsymbol{H}_{RT}$ is similar to a displacement, we will call it a *projective displacement.*

In this article we consider projective displacements which result from pure translational motions. Such a homography $\boldsymbol{H}_T$, when estimated from image measurements, contains information about the viewing geometry as well as about the observed motion. We show what information is present, where and how it is encoded, what it means geometrically, and how to extract it efficiently .

3 Algebraic Characterization

The following subsections completely characterize the structure of a projective translation in terms of algebraic similarity. Its structure is canonically described by the Jordan matrix [8]. The Jordan decompositions are given in their most general form. Finally, the fix entities of the transform are identified with those of the decomposition.

3.1 Projective Translation

The similarity class formed by the homographies of $\mathcal{P}$ which are conjugate to a rigid translation $\boldsymbol{T}_T$ up to scale is called *projective translations*. Similarity is for instance by the matrix of the Euclidean upgrade

$$\boldsymbol{H}_T = \gamma\, \boldsymbol{H}_{\mathcal{E}P}^{-1}\, \boldsymbol{T}_T\, \boldsymbol{H}_{\mathcal{E}P}, \quad \boldsymbol{T}_T = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{t} \\ \boldsymbol{0}^T & 1 \end{bmatrix}, \tag{12}$$

where γ is an unknown scale factor. Traces and determinants are invariant under similarity, so the scale γ follows from

$$4\gamma = trace\, \boldsymbol{H}_T, \ \text{ or } \gamma^4 = det\, \boldsymbol{H}_T, \tag{13}$$

Henceforth, we normalize $\boldsymbol{H}_T$ such that $\gamma = 1$.

3.2 Jordan normal form

Another similarity invariant is the characteristic polynomial $p(\lambda)$ and hence the eigenvalues λ_i. Thus $p(\lambda)$ is $(1-\lambda)^4 = 0$ and the quadruple eigenvalue $\lambda = 1$ is common to both $\boldsymbol{T}_T$ and $\boldsymbol{H}_T$. Since $rank(\boldsymbol{T}_T - \boldsymbol{I}) = 3$ and $rank(\boldsymbol{T}_T - \boldsymbol{I})^2 = 0$, the common Jordan matrix $\boldsymbol{J_T}$ of $\boldsymbol{T}_T$ and $\boldsymbol{H}_T$ has one Jordan block of order two $\boldsymbol{J}_2 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, and two of order one.

In summary, a projective translation always decomposes into its *Jordan normal form*

$$\boldsymbol{H}_T = \ \boldsymbol{H}_J^{-1}\, \boldsymbol{J}_T\, \boldsymbol{H}_J, \quad \boldsymbol{J}_T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{14}$$

(we will assume that the $\boldsymbol{J}_2$ block has been permuted into rows 3 and 4, as here).

3.3 Jordan decomposition

The similarity transform $\boldsymbol{H}_J$ in contrast is defined only up to premultiplication by any matrix $\boldsymbol{J}_C$ commuting with $\boldsymbol{J}_T\, \boldsymbol{H}_T = \left(\boldsymbol{J}_C^{-1}\boldsymbol{H}_J^{-1}\right)\boldsymbol{J}_T\, (\boldsymbol{H}_J\boldsymbol{J}_C)$, now by the similarity $\boldsymbol{H}'_J = (\boldsymbol{H}_J\boldsymbol{J}_C)$. Explicitly, the commuting class is any non-singular

$$\boldsymbol{J}_C = \mu \begin{bmatrix} j_{11} & j_{12} & 0 & j_{14} \\ j_{21} & j_{22} & 0 & j_{24} \\ j_{31} & j_{32} & 1 & j_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{15}$$

Interestingly, the third column of $\boldsymbol{H}_J$ and the fourth row of $\boldsymbol{H}_J^{-1}$ rest invariant under the application of $\boldsymbol{J}_C$, up to scales μ and $1/\mu$. Their geometric interpretation is given in section 4.1.

3.4 Eigen analysis

It is sufficient to consider the Jordan matrix, since the Eigen-spaces E_H of $\boldsymbol{H}_T$ and E_J of $\boldsymbol{J}_T$ are related by a similarity $E_H = \boldsymbol{H}_J E_J$. $E_J = \left\{[x,y,z,0]^T \mid x,y,z \in I\!\!R\right\}$ spans a hyper-plane π pointwise and its orthogonal complement $\boldsymbol{e}_J^T E_J = \boldsymbol{0}$ defines homogeneous coordinates of this plane: $\boldsymbol{e}_J = [0,0,0,1]^T$.

Conversely the dual $\hat{\boldsymbol{H}}_{\boldsymbol{T}} = \boldsymbol{H}_T^{-T}$ of a projective translation is the plane transformation

$$\hat{\boldsymbol{H}}_{\boldsymbol{T}} = \boldsymbol{H}_J^T\, \hat{\boldsymbol{J}}_{\boldsymbol{T}}\, \boldsymbol{H}_J^{-T}, \quad \hat{\boldsymbol{J}}_{\boldsymbol{T}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}. \tag{16}$$

Its Jordan matrix has the eigenspace $\hat{E}_J = [x,y,0,t]^T$. These planes intersects in the point Δ, which is the orthogonal complement $\hat{\boldsymbol{e}}_J^T\hat{E}_J = \boldsymbol{0}$ with coordinates $\hat{\boldsymbol{e}}_J = [0,0,1,0]^T$.

We finally express the fixed entities π and Δ in the original frame where $\boldsymbol{H}_T$ is represented in. It is in our case the camera frame $\mathcal{P}$. Write the columns of $\boldsymbol{H}_J^{-1}$ and the rows $\boldsymbol{H}_J$ as

$$\boldsymbol{H}_T = \begin{bmatrix} \\ \boldsymbol{c}_1 \ \boldsymbol{c}_2 \ \boldsymbol{c}_3 \ \boldsymbol{c}_4 \\ \\ \end{bmatrix} \boldsymbol{J}_T \begin{bmatrix} \boldsymbol{r}_1^T \\ \boldsymbol{r}_2^T \\ \boldsymbol{r}_3^T \\ \boldsymbol{r}_4^T \end{bmatrix}$$

$$\text{where} \boldsymbol{r}_i^T\boldsymbol{c}_j = \delta_{ij}. \tag{17}$$

$E_H = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3\}$ spans again pointwise the hyper-plane π and the orthogonal complement(17) $\boldsymbol{e}_H = \boldsymbol{r_4}$ yields the new coordinates of π . Similarily, the left eigenspace $\hat{E}_H = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_4\}$ intersects in $\Delta = \boldsymbol{c}_3$, the orthogonal completent of $\hat{\boldsymbol{e}}_H$ (17).

4 New Parameterization

We write the Jordan matrix as the 4-by-4 sum of the identity $\boldsymbol{I}$ plus the residual $\boldsymbol{E}_{34} = [1_{34}]$ (see (14)).

$$\begin{aligned} \boldsymbol{H}_T &= [\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3, \boldsymbol{c}_4]\,(\boldsymbol{E_{34}} + \boldsymbol{I})\,[\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3, \boldsymbol{r}_4]^T \\ &= \boldsymbol{H}_t + \boldsymbol{I}, \quad \boldsymbol{H}_t = \boldsymbol{c}_3\, \boldsymbol{r}_4^T, \end{aligned} \tag{18}$$

and realize that every $\boldsymbol{H}_T$ is the sum of $\boldsymbol{I}$ plus the outer product of two vectors $\boldsymbol{c}_3$ and $\boldsymbol{r}_4$, where $\boldsymbol{H}_t$ has rank 1, trace 0, and is nilpotent with order 2:

$$\boldsymbol{H}_t^2 = \boldsymbol{c}_3\, \boldsymbol{r}_4^T \boldsymbol{c}_3\, \boldsymbol{r}_4^T = \boldsymbol{c}_3\, \delta_{34}\, \boldsymbol{r}_4 = \boldsymbol{0} \tag{19}$$

Defining the scale factor $\alpha = ||\boldsymbol{c}_3||||\boldsymbol{r}_4||$ and the unit vectors $\bar{\boldsymbol{c}}_3 = \boldsymbol{c}_3/||\boldsymbol{c}_3||$ and $\bar{\boldsymbol{r}}_4 = \boldsymbol{r}_4/||\boldsymbol{r}_4||$, we have

$$\boldsymbol{H}_T = \alpha \boldsymbol{H}_{t,\alpha} + \boldsymbol{I}, \quad \boldsymbol{H}_{t,\alpha} = \bar{\boldsymbol{c}}_3\, \bar{\boldsymbol{r}}_4^T. \tag{20}$$

Formally, $\boldsymbol{c}_3$ and $\boldsymbol{r}_4$ have eight parameters and obey one constraint (17). The seven degrees of freedom are rearranged into one scalar parameter α plus two 3 dof unit vectors $\bar{\boldsymbol{c}}_3$ and $\bar{\boldsymbol{r}}_4$, in order to normally represent the underlying geometric object. We will show, that $\boldsymbol{r_4}$ is common to all projective translations observed by the same stereo system; that $\boldsymbol{c}_3$ is common to all projective translations with the same direction observed by a static stereo rig; and that α corresponds to relative distance covered in this direction.

4.1 Geometric Interpretation

Now we (12) to give a geometric interpretation of these algebraic entities. Trivially, there is an affine similarity $\boldsymbol{A}_t$ dependent on $\boldsymbol{t} = [t_x, t_y, t_z]^T$, which transforms $\boldsymbol{T}_T$ into $\boldsymbol{J}_T$. In consequence, $\boldsymbol{J}_T$ is a translation in the affine *Jordan frame* $\mathcal{J}$ induced by $\boldsymbol{A}_t$. Due to the ambiguity, there are many possible Jordan frames, however all of them are affine, since $\boldsymbol{J}_C$ is affine, too. Hence, coordinates have their standard interpretation, so $\pi = \boldsymbol{\pi}_\infty^T$ is the *plane at infinity* and $\Delta = \Delta_t$ is the *vanishing point* (Figure 1).

More rigorously, one can take

$$\boldsymbol{A}_t = \begin{bmatrix} -1/t_x & 0 & 1/t_z & 0 \\ 0 & -1/t_y & 1/t_z & 0 \\ 0 & 0 & 1/t_z & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \boldsymbol{A}_t^{-1} = \begin{bmatrix} -t_x & 0 & t_x & 0 \\ 0 & -t_y & t_x & 0 \\ 0 & 0 & t_z & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{H}_T = \boldsymbol{H}_{\mathcal{E}P}^{-1}\, \boldsymbol{A}_t^{-1}\;\; \boldsymbol{J}_T\;\; \boldsymbol{A}_t\, \boldsymbol{H}_{\mathcal{E}P} \tag{21}$$

(c.f. (12)). Then $\boldsymbol{c}_3$ and $\boldsymbol{r}_4^T$ become

$$\boldsymbol{c}_3 = \begin{bmatrix} \boldsymbol{K} & \boldsymbol{0} \\ \multicolumn{2}{c}{\cdots} \end{bmatrix} \begin{bmatrix} t_x \\ t_y \\ t_z \\ 0 \end{bmatrix}, \quad \boldsymbol{r}_4^T = \boldsymbol{a}^T, \tag{22}$$

Also remember that $\boldsymbol{c}_3$, $\boldsymbol{r}_4$ they are constant up to scale for all decompositions (section 3.3).

This shows that $\boldsymbol{r}_4$ defines the plane at infinity $\boldsymbol{\pi}_\infty^T$ in $\mathcal{P}$. It is independent of the particular translation and encodes the affine stereo geometry. Analogously $\boldsymbol{c}_3$ defines the vanishing point $\boldsymbol{\Delta}_t$ in $\mathcal{P}$. Its first three components $\boldsymbol{c}_3'$ are the the non-trivial coordinates of $\boldsymbol{\Delta}_t$ in $\mathcal{A}$, which is in fact the motion epipole in the right image. $\boldsymbol{c}_3$ depends on the translation, the intrinsic geometry, and the affine stereo geometry, but $\boldsymbol{c}_3'$ depends only on the translation. The parameter α is the relative distance covers along the translation axis, as shown in greater detail in section 5.2.

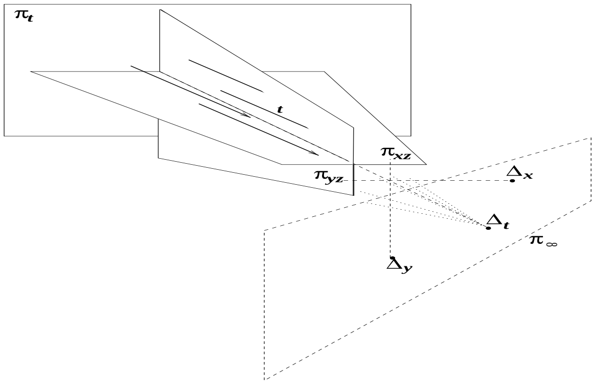

Figure 1: The point-translation's eigenspace spans the plane at infinity $\boldsymbol{\pi}_\infty^T$, e.g. by the vanishing points $\boldsymbol{\Delta}_x$, $\boldsymbol{\Delta}_y$, $\boldsymbol{\Delta}_t$. The plane-translation's eigenspace is spanned by $\boldsymbol{\pi}_\infty^T$ and two planes parallel to the translation, e.g. $\boldsymbol{\pi}_\infty^T$, $\boldsymbol{\pi}_{yz}$, $\boldsymbol{\pi}_{xz}$. There. They intersect in its vanishing point $\boldsymbol{\Delta}_t$.

5 Computational Decomposition

We aim to determine $\boldsymbol{c}_3$ and $\boldsymbol{r}_4$ from a number of consistent estimates $\boldsymbol{H}_T^{(i)}$. "Consistent" means that they need a common plane at infinity to cumulatively determine $\boldsymbol{r}_4$, and a common vanishing point to cumulatively extract $\boldsymbol{c}_3$.

The eight parameters of $\boldsymbol{c}_3$ and $\boldsymbol{r}_4$ obey 16 bilinear constraints from (18) and one from (17). Algebraic solutions are straight-forward, e.g.

$$\boldsymbol{r}_4 \simeq [\frac{\boldsymbol{H}_{T21}}{\boldsymbol{H}_{T24}}, \frac{\boldsymbol{H}_{T12}}{\boldsymbol{H}_{T14}}, \frac{\boldsymbol{H}_{T13}}{\boldsymbol{H}_{T14}}, 1]^T \tag{23}$$

$$\boldsymbol{c}_3 \simeq [\frac{\boldsymbol{H}_{T12}}{\boldsymbol{H}_{T42}}, \frac{\boldsymbol{H}_{T21}}{\boldsymbol{H}_{T41}}, \frac{\boldsymbol{H}_{T31}}{\boldsymbol{H}_{T41}}, 1]^T, \tag{24}$$

but do not fully take redundancy into account and are numerically very unstable.

The analytical Jordan decomposition (14) incorporates all of the constraints and $\boldsymbol{r}_4$ and $\boldsymbol{c}_3$ follow immediately (see section 3.3). However, the numerical detection of the locus of $\boldsymbol{J}_2$ is very unstable, as it requires the eigenvalues to be numerically equal. In both approaches, the accumulation of several $\boldsymbol{H}_T^{(i)}$ remains a problem.

5.1 Algorithm

We propose a two-step solution based on the trace and *singular value decomposition* (SVD), which is numerically stable and naturally allows the inputs to be accumulated.

Step 1) $\boldsymbol{H}_t(i)$ from $\boldsymbol{H}_T^{(i)}, i \in \{1, \dots, n\}$

- Normalize $\boldsymbol{H}_T^{(i)}$ by the trace (13) and compute $\boldsymbol{H}_t^{(i)}$ by (18).

Normalizing by the determinant or by directly calculted eigenvalues turns out to much less stable.

Step 2a) $\boldsymbol{\pi}_\infty^T$ and $\boldsymbol{\Delta}_t$ from a single $\boldsymbol{H}_t$:

- Compute the SVD, with $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \sigma_4$:

$$\boldsymbol{H}_t = \boldsymbol{U}\ diag(\sigma_1, \sigma_2, \sigma_3, \sigma_4)\ \boldsymbol{V}, \tag{25}$$

To reject outliers, check that the conditioning $\rho = \sigma_1/\sigma_2$ is large, i.e. that $\boldsymbol{H}_t$ has numerical rank 1.

$$\tilde{\boldsymbol{H}}_t = \boldsymbol{U}\ diag(\sigma_1, 0, 0, 0)\ \boldsymbol{V} = \sigma_1\ \boldsymbol{U}_{\bullet 1}\ \boldsymbol{V}_{1\bullet}$$

is the matrix of rank 1 closest to $\boldsymbol{H}_t$ with respect to matrix norms $|| \cdot ||_2$ and $|| \cdot ||_F$.

- Take the first column $\boldsymbol{U}_{\bullet 1}$=$\boldsymbol{c_3}$ as $\boldsymbol{\Delta}_t$ and the first row $\boldsymbol{V}_{1\bullet} = \boldsymbol{r}_4$ as $\boldsymbol{\pi}_\infty^T$, and obtain (18).
- Compute the unit vectors and α (20).

In contrast, algebraic solutions similar to (24) for $\boldsymbol{c}_3$ and $\boldsymbol{r}_4$ from $\boldsymbol{H}_t$ performed badly. A component-wise median filter failed completely, but the application of such robust estimators was not explored in greater depth.

Step 2b) Estimate $\boldsymbol{\pi}_\infty^T$ from accumulated $\boldsymbol{H}_T^{(i)}, i \in \{1, \dots, n\}$

- Stack all n $\boldsymbol{H}_t^{(i)}$ to

$$\boldsymbol{\mathcal{H}}^{4n \times 4} = \begin{bmatrix} \boldsymbol{H}_t^{(1)} \\ \boldsymbol{H}_t^{(2)} \\ \dots \\ \boldsymbol{H}_t^{(n)} \end{bmatrix} \tag{26}$$

- Compute the SVD $\boldsymbol{\mathcal{H}}^{4n \times 4} = \boldsymbol{UDV}$
- Take $\boldsymbol{\pi}_\infty^T$ as $\boldsymbol{r}_4 = \boldsymbol{V}_{1\bullet}$

Step 2c) Cumulative Δ_t from $\boldsymbol{H}_T^{(i)}, i \in \{1, \dots, n\}$

- Transpose and stack all n $\boldsymbol{H}_t^{(i)T}$ to

$$\boldsymbol{\mathcal{H}}^{4n \times 4} = \begin{bmatrix} \boldsymbol{H}_t^{(1)T} \\ \boldsymbol{H}_t^{(2)T} \\ \dots \\ \boldsymbol{H}_t^{(n)T} \end{bmatrix} \tag{27}$$

- Compute the SVD $\boldsymbol{\mathcal{H}}^{4n \times 4} = \boldsymbol{UDV}$
- Take $\boldsymbol{\Delta}_t$ as $\boldsymbol{c}_3 = \boldsymbol{V}_{1\bullet}$

5.2 Applications

The existance of a reliable method to estimate the motion parameters allows us to further exploit (18) or (20) in a number of applications related to translational motion. Nil-potency (19) is used often in the derivations.

- uniform, discrete extrapolation:

$$\boldsymbol{H}_T^n = \sum_{k=0}^{n} \binom{n}{k} \alpha^k \boldsymbol{H}_t^k \boldsymbol{I}^{n-k} = n\alpha \boldsymbol{H}_t + \boldsymbol{I} \tag{28}$$

From one uncalibrated observation of a translation, any integral multiple of the motion can be synthesized, and applied to a set of points ...

- exponential representation:

$$exp(\boldsymbol{H}_t) = \sum_{k=0}^{\infty} \frac{\boldsymbol{H}_t^k}{k!} = \boldsymbol{H}_t + \boldsymbol{I} = \boldsymbol{H}_T \tag{29}$$

This relationship is useful in more elaborate representation theory.

- continuous inter-/extrapolation:

$$\boldsymbol{H}_T^\alpha = exp(\alpha \boldsymbol{H}_t) = \alpha \boldsymbol{H}_t + \boldsymbol{I} \tag{30}$$

From one uncalibrated observation of a translation, any fractional translation in this direction can be synthesized, either forwards or backwards ...

- composition:

$$\begin{aligned} &\boldsymbol{H}_{TA}\boldsymbol{H}_{TB} = (\boldsymbol{H}_{ta} + \boldsymbol{I})\,(\boldsymbol{H}_{tb} + \boldsymbol{I}) \\ &= \boldsymbol{\Delta}_{ta}\boldsymbol{\pi}_\infty^T + \boldsymbol{\Delta}_{tb}\boldsymbol{\pi}_\infty^T + \boldsymbol{\Delta}_{ta}\boldsymbol{\pi}_\infty^T\boldsymbol{\Delta}_{tb}\boldsymbol{\pi}_\infty^T + \boldsymbol{I} \\ &= (\boldsymbol{\Delta}_{ta} + \boldsymbol{\Delta}_{tb})\ \boldsymbol{\pi}_\infty^T + \boldsymbol{I} \end{aligned}$$

From uncalibrated observations of two translations in different directions A and B, the resulting direction of their concatenation can be found. In fact, any linear combination of the both translations can be synthesized. Extension to n directions and resulting complete decomposition of projective translations are now straight-forward.

- affine stereo calibration

  Once $\boldsymbol{\pi}_\infty^T$ is known, projection matrices and reconstructions in the projective frame $\mathcal{P}$ can be upgraded *a-posteriori* to affine ones (8). The practical effectivness of this method is evaluation in section 6.

6 Experiments

Three different scenarios were evaluated experimentally. First, 12 translations of 18 markers on a gripper in various orientations were simulated at a distance of $90cm$ from a stereo rig with $20cm$ baseline. Second, two image sequences were taken with a standard stereo rig, which was translated along three axes. At 7 equi-distant stops per axis, stereo images of either the calibration grid or the "house-scene" were grabbed (Figure 4). A projective reconstruction was upgraded to affine using the estimated $\boldsymbol{\pi}_\infty^T$ (8). To compare this with the known Euclidean scene structure, we used two error measures:

$$\begin{aligned} e_Q &= \left| \left(\frac{(\boldsymbol{N}_2 - \boldsymbol{N}_0)}{(\boldsymbol{N}_1 - \boldsymbol{N}_0)} - \frac{(\boldsymbol{X}_2 - \boldsymbol{X}_0)}{(\boldsymbol{X}_1 - \boldsymbol{X}_0)} \right) (\boldsymbol{X}_1 - \boldsymbol{X}_0) \right| \\ e_D &= ||\boldsymbol{X} - \boldsymbol{A}_{fit}\boldsymbol{N}||, \end{aligned} \qquad (31)$$

where $(\boldsymbol{X}_2, \boldsymbol{X}_1, \boldsymbol{X}_0)$ are collinear. $\boldsymbol{A}_{fit}$ is the affinity which best fits structure in $\mathcal{A}$ to $\mathcal{E}$ in terms of Euclidean least-squares. The direct affine error e_Q scales the difference of all length-ratios to metric units. The indirect Euclidean error e_D measures how close affine structure is to the affine ground-truth that results from $\boldsymbol{A}_{fit}$. Both measures give qualitatively consistent results (Figure 2).

6.1 Gripper motion

The gripper sequence was used to study the accuracy and degeneracy of our method at increasing levels of additive Gaussian noise with σ in px(pixel). Figure 2 shows that 4 motions suffice to achieve stability. At $\sigma < 1px$, the error e_D is acceptable after only one motion and decreases below $0.2mm$. With $1 \leq \sigma \leq 2$, e_D is still below $0.5mm$, but for $\sigma \approx 3$ if $\boldsymbol{H}_T$ is estimated just linearly, the error increased rapidly.

6.2 Calibration grid

The grid sequence considers the scenario of off-line self-calibration. The image points are precise ($0.05px$), and about 100 are matched. Figure 3 shows that the results from real data are consistent with the simulations. The sections [1:5], [6:11], and [12:17] correspond to the three motion axes. The accuracy is lower when the translation is aligned with the optical axis [6:11]. The cumulative estimate is robust against derogate appositions and gains even from minor conductive appositions. The absolute error of $0.2 - 0.3mm$ compares favorably with the usual precision in structure from stereo.

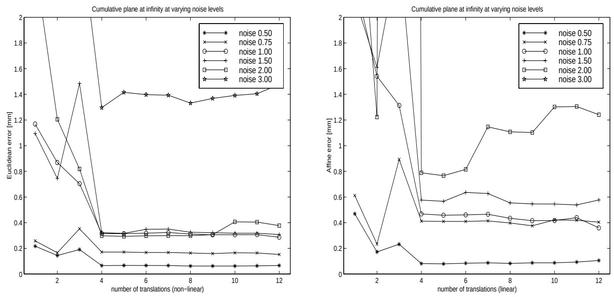

Figure 2: Accuracy at increasing noise levels quantified by e_D for non-linear $\boldsymbol{H}_T$and by e_Q for linear $\boldsymbol{H}_T$

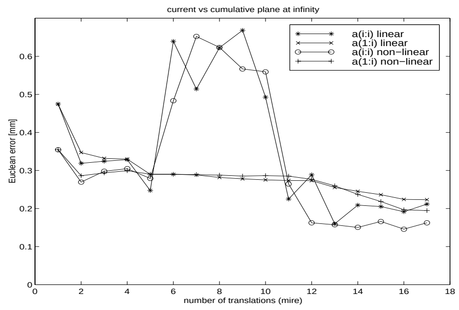

Figure 3: Comparison of single and cumulative estimates from linear and non-linear $\boldsymbol{H}_T$.

6.3 House scene

The "house sequence" considers the scenario of on-line self-calibration. The precision of matched points of interest is $1px$ and there are 1-2 false matches among a total of 30-80. The affine calibration from the house is used to upgrade a perspective reconstruction of the calibration grid (see Figure4) and is evaluated as in section 6.2. Qualitatively, stability and robustness are similar. Quantitatively, the error e_D is slightly higher but below $1mm$ as soon as $\boldsymbol{H}_T$ is estimated non-linearly.

For a scenario of visual navigation, the intrinsic parameters of motion were extracted. $\boldsymbol{H}_T$ was estimated for all 36 possible translations between 6 eqi-distant stops along one axis and decomposed following section 5.1 into (20). The error in α is roughly 0.15%, which corresponds to $0.3mm$ in space. The direction when estimated from a single motion deviates by up to $5°$ from its cumulative estimate. A comparison of the pair-wise angles between the three estimated axes with Euclidean ground-truth gave an error of $1°$ (Figure 5).

7 Summary and Conclusions

In this paper we have described a new method for the affine calibration of a stereo pair from one or more trans-

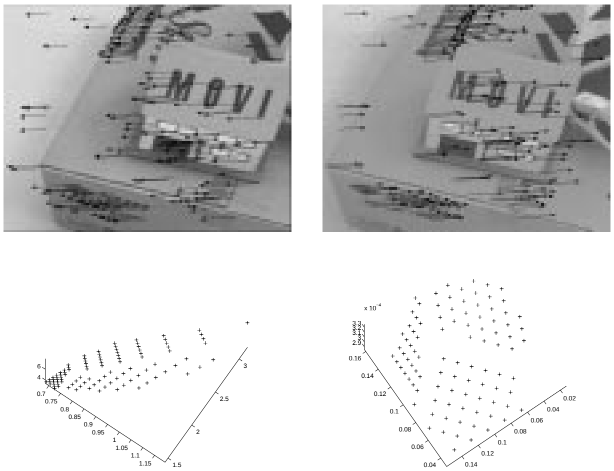

Figure 4: Motion field in house-scene and the resulting affine upgrade applied to grid.

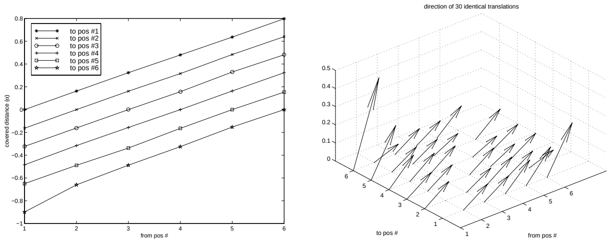

Figure 5: Distance and direction of all 36 translations.

lational motions. The method is based on an in-depth algebraic analysis of the 4×4 homography linking two projective reconstructions computed with the stereo rig before and after a translational motion. This analysis allows a simple parameterization of the homography, thus defining a *projective translation* with 7 parameters. This formulation leads to a straightforward numerical implementation within which several motions with different directions of translation can be accumulated to improve numerical stability.

The method has been applied to synthetic, calibrated and real data. In all of these cases, the method tolerated image noise with a standard deviation of up to 2 pixels provided that at least 4 motions were performed.

Recently it has been shown that, when projective structure is upgraded to affine and then to Euclidean, the affine upgrade stage is the most difficult one from a practical point of view and the final accuracy of the Euclidean reconstruction depends heavily on its accuracy [7]. Therefore we believe that the method suggested in this paper is an important contribution to the problem of self calibration of a stereo rig.

References

[1] P. A. Beardsley and A. Zisserman. Affine calibration of mobile vehicles. In *Proceedings of Europe-China Workshop on Geometrical Modelling and Invariants for Computer Vision*, pages 214–221, Xi'an, China, April 1995. Xidan University Press.

[2] F. Devernay and O. Faugeras. From projective to euclidean reconstruction. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, San Francisco, California, USA*, pages 264–269, June 1996.

[3] O. Faugeras. *Three-Dimensional Computer Vision - A Geometric Viewpoint*. Artificial intelligence. The MIT Press, Cambridge, MA, USA, Cambridge, MA, 1993.

[4] O. Faugeras. Stratification of 3-D vision: projective, affine, and metric representations. *Journal of the Optical Society of America A*, 12(3):465–484, March 1995.

[5] R. Hartley, R. Gupta, and T. Chang. Stereo from uncalibrated cameras. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Urbana-Champaign, Illinois, USA*, pages 761–764, 1992.

[6] R. Hartley and P. Sturm. Triangulation. In *Proceedings of* ARPA *Image Understanding Workshop, Monterey, California, USA*, pages 957–966, November 1994.

[7] R. Horaud and G. Csurka. Self-calibration and euclidean reconstruction using motions of a stereo rig. In *Proceedings of the 6th International Conference on Computer Vision, Bombay, India*, pages 96–103, January 1998.

[8] R. Horn and C. Johnson. *Matrix analysis*. Cambridge University Press, 1985.

[9] T. Moons, L. Van Gool, M. Proesmans, and E. Pauwels. Affine reconstruction from perspective image pairs with a relative object-camera translation in between. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(1):77–83, January 1996.

[10] L. Quan and R. Mohr. Affine shape representation from motion through reference points. *Journal of Mathematical Imaging and Vision*, 1:145–151, 1992.

[11] A. Zisserman, P. Beardsley, and I. Reid. Metric calibration of a stereo rig. In *Workshop on Representation of Visual Scenes, Cambridge, Massachusetts, USA*, pages 93–100, June 1995.