Structure and Motion Estimation using Sparse Point and Line Correspondences in Multiple Affine Views[*]

Lars Bretzner and Tony Lindeberg

Computational Vision and Active Perception Laboratory (CVAP)
Department of Numerical Analysis and Computing Science
KTH (Royal Institute of Technology)
S-100 44 Stockholm, Sweden.

http://www.nada.kth.se/˜tony
Email: { bretzner, tony}@nada.kth.se

Technical report ISRN KTH/NA/P–99/13–SE

Abstract

This paper addresses the problem of computing three-dimensional structure and motion from an unknown rigid configuration of points and lines viewed by an affine projection model. An algebraic structure, analogous to the trilinear tensor for three perspective cameras, is defined for configurations of three centered affine cameras. This centered affine trifocal tensor contains 12 non-zero coefficients and involves linear relations between point correspondences and trilinear relations between line correspondences. It is shown how the affine trifocal tensor relates to the perspective trilinear tensor, and how three-dimensional motion can be computed from this tensor in a straightforward manner. A factorization approach is developed to handle point features and line features simultaneously in image sequences, and degenerate feature configurations are analysed. This theory is applied to a specific problem in human-computer interaction of capturing three-dimensional rotations from gestures of a human hand. This application to quantitative gesture analyses illustrates the usefulness of the affine trifocal tensor in a situation where sufficient information is not available to compute the perspective trilinear tensor, while the geometry requires point correspondences as well as line correspondences over at least three views.

[*] An earlier version of this manuscript was presented in H. Burkhardt and B. Neumann (eds.) *Proc. 5th European Conference on Computer Vision*, (Freiburg, Germany), vol. 1406 of Springer-Verlag Lecture Notes in Computer Science, pp. 141–157, June 1998. The support from the Swedish Research Council for Engineering Sciences, TFR, and the Swedish National Board for Industrial and Technical Development, NUTEK, is gratefully acknowledged.

Contents

1 Introduction

The problem of deriving structural information and motion cues from image sequences arises as an important subproblem in several computer vision tasks. In this paper, we are concerned with the computation of three-dimensional structure and motion from point and line correspondences extracted from a rigid three-dimensional object of unknown shape, using the affine camera model.

Early works addressing this problem domain based on point correspondences from perspective and orthographic projection have been presented by Ullman [1], Maybank [2], Huang and Lee [3], Huang and Netravali [4] and others. With the introduction of the affine camera model (Koenderink and van Doorn [5], Mundy and Zisserman [6]) a large number of approaches have been developed, including (Shapiro [7], Beardsley et al. [8], McLauchlan et al. [9], Torr [10]) to mention just a few, see also (Faugeras [11]). Line correspondences have been studied by (Spetsakis and Aloimonos [12], Weng et al [13]), and factorization methods for points and lines constitute a particularly interesting development (Tomasi and Kanade [14], Poelman and Kanade [15], Quan and Kanade [16], Sturm and Triggs [17]). These directions of research have recently been combined with the ideas behind the fundamental matrix (Longuet-Higgins [18], Faugeras [19], Xu and Zhang [20]) and have lead to the trilinear tensor (Shashua [21], Hartley [22], Heyden [23]) as a unified model for point and line correspondences for three cameras, with interesting applications (Beardsley et al. [24]) as well as a deeper understanding of the relations between point features and line features over multiple views (Faugeras and Mourrain [25], Heyden et al. [26]).

The subject of this paper is to build upon the abovementioned works, and to develop a framework for handling point and line features simultaneously for three or more affine views. Initially, we shall focus on image triplets and show how an *affine trifocal tensor* can be defined for three centered affine cameras. This tensor has a similar algebraic structure as the trilinear tensor for three perspective cameras. Compared to the trilinear tensor, however, it has the advantage that it contains a smaller number of coefficients, which implies that fewer feature correspondences are required to determine this tensor. Motion estimation from this tensor is more straightforward than for the perspective trilinear tensor. Moreover, the results from affine motion estimation can be expected to be more robust than perspective analysis in situations when the perspective effects are small. The handle image features in more than three images, we shall also develop a factorization approach, which involves simultaneous handling of point and line features in multiple image frames.

This theory will then be applied to the problem of computing changes in three-dimensional orientation from a sparse set of point and line correspondences. Specifically, it will be demonstrated how a man-machine interface for 3-D interaction can be designed based on the theory presented. The idea is to track point and line features corresponding to the finger tips and the orientation of the fingers, and to compute three-dimensional rotations (and translations) assuming rigidity of the hand. These motion estimates can then be used for controlling the motion of other computer-controlled equipment (Lindeberg and Bretzner [27]). Notably, we thereby eliminate the need for other external control equipment than the operator's own hand.

2 Geometric problem and extraction of image features

A main rationale to this work originates from the following question: If we have a sparse set of image features that have been tracked over a relatively long period of time, to what extent can such extended feature trajectories be used for computing the three-dimensional structure and motion of a rigid object? Moreover, we are interested in exploring whether it is possible to make use of image features that have been extracted from natural objects. Most works on three-dimensional structure and motion estimation have been performed under different conditions, by exploiting dense sets of image features, which have been computed from man-made objects.

Figure 1 shows one specific application, which we will focus on. The idea is to capture three-dimensional motions as mediated by the gestures of a human hand, and to use measurements of 3-D rotational information computed in this way for controlling other computerized equipment, see [27] for a more general description and Cipolla et al. [28], Freeman and Weissman [29], Maggioni and Kämmerer [30] for related works. In contrast to previous approaches for human–computer interaction that are based on detailed geometric hand models (such as Kuch and Huang [31], Lee and Kunii [32], Heap and Hogg [33], Yasumuro et al. [34]), we shall here explore a model based on qualitative features only. This model involves three to five fingers, and for each finger the position of the finger tip and the orientation of the finger are measured in the image domain. Successful tracking of these image features over time leads to a set of point correspondences and line correspondences. The task is then to compute changes in the 3-D orientation of such a configuration, which is assumed to be rigid.

Given only a a small number of image features, neither the trajectories of the point features or the line features *per se* are sufficient to compute the motion information we are interested in. For example, when a user holds his hand with the fingers spreading out, we have experienced that the positions of the finger tips will often be in approximately the same plane, leading to ill-conditioned motion estimates if computed from point features only. Therefore, the ability to combine point features and line features is of high importance. Moreover, due to the small number of image features, the information is not sufficient to compute the trilinear tensor for perspective projection (see the next section). For this reason, we shall use an affine projection model, and the affine trifocal tensor will be a key tool.

The trajectories of image features used as input are extracted using a framework for feature tracking with automatic scale selection reported in (Bretzner and Lindeberg [35, 36]). Blob features corresponding to the finger tips are computed from points $(x, y;\ t)$ in scale-space (Koenderink [37], Lindeberg [38]) at which the squared normalized Laplacian

$$(\nabla^2_{norm} L)^2 = t^2\,(L_{xx} + L_{yy})^2 \tag{1}$$

assumes maxima with respect to scale and space simultaneously (Lindeberg [39]). Such points are referred to as scale-space maxima of the normalized Laplacian. In a similar way, ridge features are detected from scale-space maxima of a normalized measure of ridge strength

$$\mathcal{A}L^2_{\gamma-norm} \;=\; t^{4\gamma}\,(L^2_{pp} - L^2_{qq})^2 = t^{4\gamma}\,\left((L_{xx} - L_{yy})^2 + 4L^2_{xy}\right)^2, \tag{2}$$

where L_{pp} and L_{qq} are the eigenvalues of the Hessian matrix and the normalization parameter $\gamma = 0.875$ (Lindeberg [40]). At each ridge feature, a windowed second moment matrix

$$\mu = \int\int_{(\xi,\eta)\in\mathbb{R}^2} \begin{pmatrix} L_x^2 & L_xL_y \\ L_xL_y & L_y^2 \end{pmatrix} g(\xi,\eta;\ s)\, d\xi\, d\eta \tag{3}$$

is computed using a Gaussian window function $g(\cdot,\cdot;\ s)$ centered at the spatial maximum of $\mathcal{A}L_{\gamma-norm}$ and with the integration scale s tuned by the detection scale of the scale-space maximum of $\mathcal{A}L_{\gamma-norm}$. The eigenvector of μ corresponding to the largest eigenvalue gives the orientation of the finger.

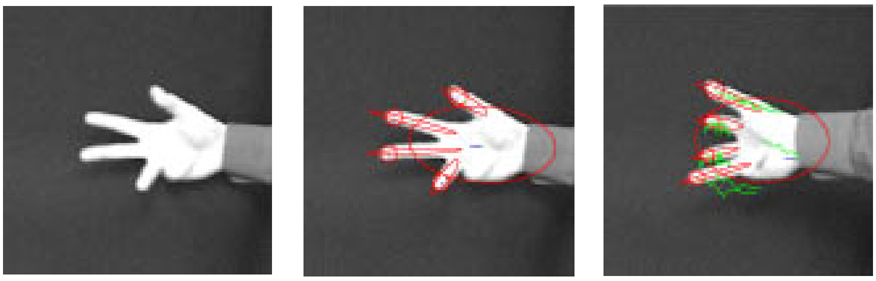


Figure 1: Results of multi-scale tracking of point and line features corresponding to the finger tips and the fingers of a human hand. (left) grey-level image showing the first frame in an image sequence, (middle) image features extracted by combining the detection of scale-space maxima of blob and ridge features [39, 40] with a qualitative hand model in the form of a multi-scale feature hierarchy [41], (right) feature trajectories obtained by multi-scale feature tracking [35].

Figure 1(c) shows an example of image trajectories obtained in this way. An attractive property of this feature tracking scheme is that the scale selection mechanism adapts the scale levels to the local image structure. This gives the ability to track image features over large size variations, which is particularly important for the ridge tracker. Provided that the contrast to the background is sufficient, this scheme gives feature trajectories over large numbers of frames, using a conceptually very simple interframe matching mechanism.

3 The trifocal tensor for three centered affine cameras

To capture motion information from the projections of an unknown configuration of points and lines in 3-D, it is necessary to have at least three independent views. A canonical model for describing the geometric relationships between point correspondences and line correspondences over three perspective views is provided by the trilinear tensor (Shashua [21, 42], Hartley [22], Heyden et al. [26]). For affine cameras, a compact model of point correspondences over multiple frames can be obtained by factorizing a matrix with image measurements to the product of two matrices of rank 3, one representing motion, and the other one representing shape (Tomasi and Kanade [14], Ullman and Basri [43]). Frameworks for capturing line correspondences over multiple affine views have been presented by Quan and Kanade [16] and for point features under perspective projection by Sturm and Triggs [17].

The subject of this section is to combine the idea behind the trilinear tensor for simultaneous modelling of point and line correspondences over three views with the affine projection model. It will be shown how an algebraic structure closely related to the trilinear tensor can be defined for three centered affine cameras. This *centered affine trifocal tensor* involves linear relations between the point features and trilinear relationships between the line features.

3.1 Perspective camera and three views

Consider a point $P = (x, y, 1, \lambda)^T$ which is projected by three camera matrices $M = [I, 0]$, $M' = [A, u']$ and $M'' = [B, u'']$ to the image points p, p' and p'':

$$p = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \\ \lambda \end{pmatrix}, \tag{4}$$

$$p' = \alpha \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} a_1^1 & a_2^1 & a_3^1 & u'^1 \\ a_1^2 & a_2^2 & a_3^2 & u'^2 \\ a_1^3 & a_2^3 & a_3^3 & u'^3 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \\ \lambda \end{pmatrix} = \begin{pmatrix} {a^1}^T p + \lambda u'^1 \\ {a^2}^T p + \lambda u'^2 \\ {a^3}^T p + \lambda u'^3 \end{pmatrix}, \tag{5}$$

$$p'' = \beta \begin{pmatrix} x'' \\ y'' \\ 1 \end{pmatrix} = \begin{pmatrix} b_1^1 & b_2^1 & b_3^1 & u''^1 \\ b_1^2 & b_2^2 & b_3^2 & u''^2 \\ b_1^3 & b_2^3 & b_3^3 & u''^3 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \\ \lambda \end{pmatrix} = \begin{pmatrix} {b^1}^T p + \lambda u''^1 \\ {b^2}^T p + \lambda u''^2 \\ {b^3}^T p + \lambda u''^3 \end{pmatrix}. \tag{6}$$

Following Faugeras and Mourrain [25] and Shashua [42], introduce the following two matrices

$$r_j^\mu = \begin{pmatrix} -1 & 0 & x' \\ 0 & -1 & y' \end{pmatrix}, \qquad s_k^\nu = \begin{pmatrix} -1 & 0 & x'' \\ 0 & -1 & y'' \end{pmatrix}. \tag{7}$$

Then, in terms of tensor notation (where $i, j, k \in [1, 3]$, $\mu, \nu \in [1, 2]$ and we follow the Einstein summation convention that a double occurrence of an index implies summation over that index) the relations between the image coordinates and the camera geometry can be written

$$\lambda r_j^\mu u'^j + r_j^\mu a_i^j p^i = 0, \qquad \lambda s_k^\nu u''^k + s_k^\nu b_i^k p^i = 0. \tag{8}$$

By introducing the trifocal tensor (Shashua [21], Hartley [22])

$$T_i^{jk} = a_i^j u''^k - b_i^k u'^j, \tag{9}$$

the relations between the point correspondences lead to the trifocal constraint

$$r_j^\mu \, s_k^\nu \, T_i^{jk} = 0. \tag{10}$$

Written out explicitly, this expression corresponds to the following four (independent) relations between the projections p, p' and p'' of P (Shashua [42]):

$$\begin{aligned} x'' T_i^{13} p^i - x'' x' T_i^{33} p^i + x' T_i^{31} p^i - T_i^{11} p^i &= 0, \\ y'' T_i^{13} p^i - y'' x' T_i^{33} p^i + x' T_i^{32} p^i - T_i^{12} p^i &= 0, \\ x'' T_i^{23} p^i - x'' y' T_i^{33} p^i + y' T_i^{31} p^i - T_i^{21} p^i &= 0, \\ y'' T_i^{23} p^i - y'' y' T_i^{33} p^i + y' T_i^{32} p^i - T_i^{22} p^i &= 0. \end{aligned} \tag{11}$$

Given three corresponding lines, $l^T p = 0$, $l'^T p' = 0$ and $l''^T p'' = 0$, each image line defines a plane through the center of projection, given by $L^T P = 0$, $L'^T P = 0$ and $L''^T P = 0$, where

$$\begin{aligned} L^T &= l^T M = (l_1,\, l_2,\, l_3\, 0), \\ L'^T &= l'^T M' = (l'_j\, a^j_1,\, l'_j\, a^j_2,\, l'_j\, a^j_3,\, l'_j\, u'^j), \\ L''^T &= l''^T M'' = (l''_k\, b^k_1,\, l''_k\, b^k_2,\, l''_k\, b^k_3,\, l''_k\, u''^k). \end{aligned} \tag{12}$$

Since l, l' and l'' are assumed to be projections of the same three-dimensional line, the intersection of the planes L, L' and L'' must degenerate to a line and

$$\text{rank} \begin{pmatrix} l_1 & l'_j\, a^j_1 & l''_k\, b^k_1 \\ l_2 & l'_j\, a^j_2 & l''_k\, b^k_2 \\ l_3 & l'_j\, a^j_3 & l''_k\, b^k_3 \\ 0 & l'_j\, u'^j & l''_k\, u''^k \end{pmatrix} = 2. \tag{13}$$

All 3×3 minors must be zero, and removal of the three first lines respectively, leads to the following trilinear relationships, out of which two are independent:

$$\begin{aligned} (l_2 T^{jk}_3 - l_3 T^{jk}_2)\, l'_j\, l''_k &= 0, \\ (l_1 T^{jk}_3 - l_3 T^{jk}_1)\, l'_j\, l''_k &= 0, \\ (l_1 T^{jk}_2 - l_2 T^{jk}_1)\, l'_j\, l''_k &= 0. \end{aligned} \tag{14}$$

These expressions provide a compact characterization of the trilinear line relations first introduced by Spetsakis and Aloimonos [12].

In summary, each point correspondence gives four equations, and each line correspondence two. Hence, K points and L lines are (generically) sufficient to express a linear algorithm for computing the trilinear tensor (up to scale) if $4K + 2L \geq 26$ (Shashua [21], Hartley [22]).

3.2 Affine camera and three views

Consider next a point $Q = (x, y, \lambda, 1)^T$ which is projected to the image points q, q' and q'' by three affine camera matrices M, M' and M'', respectively:

$$q = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = MQ = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ \lambda \\ 1 \end{pmatrix} \tag{15}$$

$$q' = \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = M'Q = \begin{pmatrix} c^1_1 & c^1_2 & c^1_3 & v'^1 \\ c^2_1 & c^2_2 & c^2_3 & v'^2 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ \lambda \\ 1 \end{pmatrix} \tag{16}$$

$$q'' = \begin{pmatrix} x'' \\ y'' \\ 1 \end{pmatrix} = M''Q = \begin{pmatrix} d^1_1 & d^1_2 & d^1_3 & v''^1 \\ d^2_1 & d^2_2 & d^2_3 & v''^2 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ \lambda \\ 1 \end{pmatrix} \tag{17}$$

Here, the parameterization of Q differs from P, since for an image point $q = (x, y, 1)^T$ the projection (15) implies that the three-dimensional point is on the ray $Q = (x, y, \lambda, 1)^T$ for some λ. By eliminating λ, we obtain the following linear relationships:

$$\begin{aligned}
(c_3^1 d_1^1 - c_1^1 d_3^1)x + (c_3^1 d_2^1 - c_2^1 d_3^1)y + d_3^1 x' - c_3^1 x'' + (c_3^1 v''^1 - d_3^1 v'^1) &= 0,\\
(c_3^2 d_1^1 - c_1^2 d_3^1)x + (c_3^2 d_2^1 - c_2^2 d_3^1)y + d_3^1 y' - c_3^2 x'' + (c_3^2 v''^1 - d_3^1 v'^2) &= 0,\\
(c_3^1 d_1^2 - c_1^1 d_3^2)x + (c_3^1 d_2^2 - c_2^1 d_3^2)y + d_3^2 x' - c_3^1 y'' + (c_3^2 v''^2 - d_3^2 v'^2) &= 0,\\
(c_3^2 d_1^2 - c_1^2 d_3^2)x + (c_3^2 d_2^2 - c_2^2 d_3^2)y + d_3^2 y' - c_3^2 y'' + (c_3^2 v''^2 - d_3^2 v'^2) &= 0.
\end{aligned} \tag{18}$$

This structure corresponds to the trilinear constraint (11) for perspective projection, and we shall refer to it as the affine trifocal point constraint.

Three lines $l^T q = 0$, $l'^T q' = 0$ and $l''^T q'' = 0$ in the three images define three planes $L^T Q = 0$, $L'^T Q = 0$ and $L''^T Q = 0$ in three-dimensional space with

$$\begin{aligned}
L^T &= l^T M = (l_1,\, l_2,\, 0,\, l_3),\\
L'^T &= l'^T M' = (l'_1 c_1^1 + l'_2 c_1^2,\, l'_1 c_2^1 + l'_2 c_2^2,\, l'_1 c_3^1 + l'_2 c_3^2,\, l'_1 v'^1 + l'_2 v'^2 + l'_3),\\
L''^T &= l''^T M'' = (l''_1 d_1^1 + l''_2 d_1^2,\, l''_1 d_2^1 + l''_2 d_2^2,\, l''_1 d_3^1 + l''_2 d_3^2,\, l''_1 v''^1 + l''_2 v''^2 + l''_3).
\end{aligned}$$

Since l, l' and l'' are projections of the same three-dimensional line, the intersection of L, L' and L'' must degenerate to a line and

$$\operatorname{rank}\begin{pmatrix}
l_1 & l'_1 c_1^1 + l'_2 c_1^2 & l''_1 d_1^1 + l''_2 d_1^2\\
l_2 & l'_1 c_2^1 + l'_2 c_2^2 & l''_1 d_2^1 + l''_2 d_2^2\\
0 & l'_1 c_3^1 + l'_2 c_3^2 & l''_1 d_3^1 + l''_2 d_3^2\\
l_3 & l'_1 v'^1 + l'_2 v'^2 + l'_3 & l''_1 v'^1 + l''_2 v''^2 + l''_3
\end{pmatrix} = 2. \tag{19}$$

All 3×3 minors must be zero, and deletion of the first, second and fourth rows, respectively, results in the following relationships between l, l' and l'':

$$\begin{aligned}
l_2\,(c_3^j v''^k - d_3^k v'^j)\,l'_j l''_k - l_3\,(c_3^j d_2^k - c_2^j d_3^k)\,l'_j l''_k &= 0,\\
l_1\,(c_3^j v''^k - d_3^k v'^j)\,l'_j l''_k - l_3\,(c_3^j d_1^k - c_1^j d_3^k)\,l'_j l''_k &= 0,\\
l_1\,(c_2^j d_3^k - c_3^j d_2^k)\,l'_j l''_k - l_2\,(c_1^j d_3^k - c_3^j d_1^k)\,l'_j l''_k &= 0,
\end{aligned} \tag{20}$$

where $c_j^3 = d_k^3 = 0$, $v'^3 = v''^3 = 1$ and only two of the relations are independent. This treatment, which largely derives similar results as Torr [10] while using another formalism, shows that point and line correspondences are captured by 16 coefficients. Each point correspondence gives four equations, and each line correspondence two. Thus, K point correspondences and L line correspondences are sufficient to compute this *affine trifocal tensor* (up to scale) if $4K + 2L \geq 15$.

4 The centered affine camera and its relations to perspective

Structurally, there is a strong similarity between the relationships for the affine camera and the corresponding relationships (11) and (14) for the perspective camera. Let us make the following formal replacements between the affine camera model (15)–(17) and the perspective camera model (4)–(6):

- Interchange rows 3 and 4 in the coordinate vectors in the 3-D domain:

$$P = (x, y, 1, \lambda)^T \quad \Rightarrow \quad Q = (x, y, \lambda, 1)^T, \tag{21}$$

- Interchange columns 3 and 4 in the camera matrices:

$$\begin{pmatrix} a_1^1 & a_2^1 & a_3^1 & a_4^1 \\ a_1^2 & a_2^2 & a_3^2 & a_4^2 \\ a_1^3 & a_2^3 & a_3^3 & a_4^3 \end{pmatrix} = \begin{pmatrix} a_1^1 & a_2^1 & a_3^1 & u'^1 \\ a_1^2 & a_2^2 & a_3^2 & u'^2 \\ a_1^3 & a_2^3 & a_3^3 & u'^3 \end{pmatrix} = \begin{pmatrix} c_1^1 & c_2^1 & v'^1 & c_3^1 \\ c_1^2 & c_2^2 & v'^2 & c_3^2 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$\begin{pmatrix} b_1^1 & b_2^1 & b_3^1 & b_4^1 \\ b_1^2 & b_2^2 & b_3^2 & b_4^2 \\ b_1^3 & b_2^3 & b_3^3 & b_4^3 \end{pmatrix} = \begin{pmatrix} b_1^1 & b_2^1 & b_3^1 & u''^1 \\ b_1^2 & b_2^2 & b_3^2 & u''^2 \\ b_1^3 & b_2^3 & b_3^3 & u''^3 \end{pmatrix} = \begin{pmatrix} d_1^1 & d_2^1 & v''^1 & d_3^1 \\ d_1^2 & d_2^2 & v''^2 & d_3^2 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{22}$$

Then, the algebraic structure between corresponding points and lines will be the same for the two projection models. This implies that the relations between point and line correspondences for three affine cameras can be expressed on the form (11) and (14) with the *affine trifocal tensor* defined by

$$\mathcal{T}_i^{jk} = a_i^j b_4^k - b_i^k a_4^j = a_i^j u''^k - b_i^k u'^j = \{(22)\} = c_i^j d_3^k - d_i^k c_3^j \tag{23}$$

where a and b are defined as in (22).

Let us next consider the case when image coordinates in the affine camera are measured relative to the center of gravity of a point configuration. This centered affine camera is obtained by setting $(v'^1, v'^2) = (v''^1, v''^2) = (0, 0)$ in (16) and (17) and corresponds to disregarding the translational motion. Written out explicitly, the components of the corresponding *centered affine trifocal tensor* $\mathcal{T}_i^{jk}$ are given by

$$\begin{array}{lll}
\mathcal{T}_1^{11} = c_1^1 d_3^1 - d_1^1 c_3^1, & \mathcal{T}_1^{12} = c_1^1 d_3^2 - d_1^2 c_3^1, & \mathcal{T}_1^{13} = c_1^1 d_3^3 - d_1^3 c_3^1 \quad = 0, \\
\mathcal{T}_1^{21} = c_1^2 d_3^1 - d_1^1 c_3^2, & \mathcal{T}_1^{22} = c_1^2 d_3^2 - d_1^2 c_3^2, & \mathcal{T}_1^{23} = c_1^2 d_3^3 - d_1^3 c_3^2 \quad = 0, \\
\mathcal{T}_1^{31} = c_1^3 d_3^1 - d_1^1 c_3^3 \quad = 0, & \mathcal{T}_1^{32} = c_1^3 d_3^2 - d_1^2 c_3^3 \quad = 0, & \mathcal{T}_1^{33} = c_1^3 d_3^3 - d_1^3 c_3^3 \quad = 0, \\
\mathcal{T}_2^{11} = c_2^1 d_3^1 - d_2^1 c_3^1, & \mathcal{T}_2^{12} = c_2^1 d_3^2 - d_2^2 c_3^1, & \mathcal{T}_2^{13} = c_2^1 d_3^3 - d_2^3 c_3^1 \quad = 0, \\
\mathcal{T}_2^{21} = c_2^2 d_3^1 - d_2^1 c_3^2, & \mathcal{T}_2^{22} = c_2^2 d_3^2 - d_2^2 c_3^2, & \mathcal{T}_2^{23} = c_2^2 d_3^3 - d_2^3 c_3^2 \quad = 0, \\
\mathcal{T}_2^{31} = c_2^3 d_3^1 - d_2^1 c_3^3 \quad = 0, & \mathcal{T}_2^{32} = c_2^3 d_3^2 - d_2^2 c_3^3 \quad = 0, & \mathcal{T}_2^{33} = c_2^3 d_3^3 - d_2^3 c_3^3 \quad = 0, \\
\mathcal{T}_3^{11} = v'^1 d_3^1 - v''^1 c_3^1 = 0, & \mathcal{T}_3^{12} = v'^1 d_3^2 - v''^2 c_3^1 = 0, & \mathcal{T}_3^{13} = v'^1 d_3^3 - v''^3 c_3^1 = -c_3^1, \\
\mathcal{T}_3^{21} = v'^2 d_3^1 - v''^1 c_3^2 = 0, & \mathcal{T}_3^{22} = v'^2 d_3^2 - v''^2 c_3^2 = 0, & \mathcal{T}_3^{23} = v'^2 d_3^3 - v''^3 c_3^2 = -c_3^2, \\
\mathcal{T}_3^{31} = v'^3 d_3^1 - v''^1 c_3^3 = d_3^1, & \mathcal{T}_3^{32} = v'^3 d_3^2 - v''^2 c_3^3 = d_3^2, & \mathcal{T}_3^{33} = v'^3 d_3^3 - v''^3 c_3^3 \quad = 0,
\end{array} \tag{24}$$

and the relations between point and line correspondences in (18) and (20) can be written

$$\begin{aligned}
\mathcal{T}_3^{13} x'' + \mathcal{T}_3^{31} x' - \mathcal{T}_1^{11} x - \mathcal{T}_2^{11} y &= 0, \\
\mathcal{T}_3^{13} y'' + \mathcal{T}_3^{32} x' - \mathcal{T}_1^{12} x - \mathcal{T}_2^{12} y &= 0, \\
\mathcal{T}_3^{23} x'' + \mathcal{T}_3^{31} y' - \mathcal{T}_1^{21} x - \mathcal{T}_2^{21} y &= 0, \\
\mathcal{T}_3^{23} y'' + \mathcal{T}_3^{32} y' - \mathcal{T}_1^{22} x - \mathcal{T}_2^{22} y &= 0,
\end{aligned} \tag{25}$$

$$
\begin{aligned}
&l_3(l'_1l''_1\mathcal{T}_1^{11}+l'_1l''_2\mathcal{T}_1^{12}+l'_2l''_1\mathcal{T}_1^{21}+l'_2l''_2\mathcal{T}_1^{22})-l_1(l'_1l''_3\mathcal{T}_3^{13}+l'_2l''_3\mathcal{T}_3^{23}+l'_3l''_1\mathcal{T}_3^{31}+l'_3l''_2\mathcal{T}_3^{32})=0,\\
&l_3(l'_1l''_1\mathcal{T}_2^{11}+l'_1l''_2\mathcal{T}_2^{12}+l'_2l''_1\mathcal{T}_2^{21}+l'_2l''_2\mathcal{T}_2^{22})-l_2(l'_1l''_3\mathcal{T}_3^{13}+l'_2l''_3\mathcal{T}_3^{23}+l'_3l''_1\mathcal{T}_3^{31}+l'_3l''_2\mathcal{T}_3^{32})=0,\\
&l_1(l'_1l''_1\mathcal{T}_2^{11}+l'_1l''_2\mathcal{T}_2^{12}+l'_2l''_1\mathcal{T}_2^{21}+l'_2l''_2\mathcal{T}_2^{22})-l_2(l'_1l''_1\mathcal{T}_1^{11}+l'_1l''_2\mathcal{T}_1^{12}+l'_2l''_1\mathcal{T}_1^{21}+l'_2l''_2\mathcal{T}_1^{22})=0.
\end{aligned}
\tag{26}
$$

In the generic case, only two of the three relations in (26) are independent. The third relation is, however, needed, since the first two relations vanish when $l_3 = l'_3 = l''_3 = 0$, *i.e.*, when the 3D line goes through the origin of the world coordinate system (in our case is the center of gravity of the 3D point configuration).

The centered affine trifocal tensor has 12 non-zero entries. Due to the centering of the equations, one point correspondence is redundant. Thus, K point correspondences and L line correspondences are (generically) sufficient to compute $\mathcal{T}_i^{jk}$ (up to scale) provided that $4(K-1)+2L \geq 11$.

5 Orientation from the centered affine trifocal tensor

To compute the camera parameters from the affine trifocal tensor, we largely follow the approach that Hartley [22] uses for three perspective cameras. The calculations can, however, be simplified with affine cameras. From (24) we directly get

$$c_3^1 = -\mathcal{T}_3^{13}, \qquad d_3^1 = \mathcal{T}_3^{31}, \qquad c_3^2 = -\mathcal{T}_3^{23}, \qquad d_3^2 = \mathcal{T}_3^{32}. \tag{27}$$

Given these c_3^j and d_3^k, the remaining c_i^j and d_i^k can be computed from (24) using

$$
\begin{pmatrix}
d_3^1 & & & & -c_3^1 & & & \\
d_3^2 & & & & & -c_3^1 & & \\
 & d_3^1 & & & -c_3^2 & & & \\
 & d_3^2 & & & & -c_3^2 & & \\
 & & d_3^1 & & & & -c_3^1 & \\
 & & d_3^2 & & & & & -c_3^1 \\
 & & & d_3^1 & & & -c_3^2 & \\
 & & & d_3^2 & & & & -c_3^2
\end{pmatrix}
\begin{pmatrix} c_1^1 \\ c_1^2 \\ c_2^1 \\ c_2^2 \\ d_1^1 \\ d_1^2 \\ d_2^1 \\ d_2^2 \end{pmatrix}
=
\begin{pmatrix} \mathcal{T}_1^{11} \\ \mathcal{T}_1^{12} \\ \mathcal{T}_1^{21} \\ \mathcal{T}_1^{22} \\ \mathcal{T}_2^{11} \\ \mathcal{T}_2^{12} \\ \mathcal{T}_2^{21} \\ \mathcal{T}_2^{22} \end{pmatrix}. \tag{28}
$$

The camera matrices are, however, not uniquely determined. The centered affine trifocal tensor $\mathcal{T}_i^{jk}$ in (23) is invariant under transformations of the type $\tilde{c}_i^j = c_i^j + \gamma_i\, c_3^j$ and $\tilde{d}_i^k = d_i^k + \gamma_i\, d_3^k$.

With N' and N'' denoting the upper left 2×3 submatrices of M' and M'' respectively, this ambiguity implies that both $\{\tilde{N}', \tilde{N}''\}$ and $\{N', N''\}$ are possible solutions (with $\tilde{N}''$ analogously)

$$\tilde{N}' = \begin{pmatrix} \tilde{c}_1^1 & \tilde{c}_2^1 & \tilde{c}_3^1 \\ \tilde{c}_1^2 & \tilde{c}_2^2 & \tilde{c}_3^2 \end{pmatrix} = \begin{pmatrix} c_1^1 & c_2^1 & c_3^1 \\ c_1^2 & c_2^2 & c_3^2 \end{pmatrix} \begin{pmatrix} 1 & & \\ & 1 & \\ \gamma_1 & \gamma_2 & \gamma_3 \end{pmatrix} = N'\,\Gamma. \tag{29}$$

To determine Γ, let us assume that the affine camera model corresponds to scaled orthographic projection, and that internal calibration is available. Then, the camera matrices can be written (with $\tilde{N}''$ analogously)

$$\tilde{N}' = \sigma' \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} R' = \sigma' \begin{pmatrix} \rho'^1_1 & \rho'^1_2 & \rho'^1_3 \\ \rho'^2_1 & \rho'^2_2 & \rho'^2_3 \end{pmatrix}, \tag{30}$$

where ${\rho'^j}^T = ({\rho'}^j_1, {\rho'}^j_2, {\rho'}^j_3)$ are the row vectors in the three-dimensional rotation matrix R', while σ' is a scaling factor. Since the rows of R' are orthogonal, ${\rho'^j_i}^T {\rho'}^k_i = \delta^{jk}$, where δ^{jk} is the Kronecker delta symbol, we have

$$\tilde{N}'{\tilde{N}'}^T = N'\Gamma\Gamma^T {N'}^T = (\sigma')^2 I_{2\times 2}, \tag{31}$$

where $I_{2\times 2}$ represents a unit matrix of size 2 × 2. With

$$\Gamma\Gamma^T = \begin{pmatrix} 1 & 0 & \gamma_1 \\ 1 & 0 & \gamma_2 \\ \gamma_1 & \gamma_2 & \gamma_1^2+\gamma_2^2+\gamma_3^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \xi \\ 0 & 1 & \eta \\ \xi & \eta & \zeta \end{pmatrix} \tag{32}$$

we rewrite (31) as

$$\begin{pmatrix} 2c_1^1c_3^1 & 2c_2^1c_3^1 & (c_3^1)^2 & -1 & 0 \\ c_1^1c_3^2 + c_3^1c_1^2 & c_2^1c_3^2 + c_3^1c_2^2 & c_3^1c_3^2 & 0 & 0 \\ 2c_1^2c_3^2 & 2c_2^2c_3^2 & (c_3^2)^2 & -1 & 0 \\ 2d_1^1d_3^1 & 2d_2^1d_3^1 & (d_3^1)^2 & 0 & -1 \\ d_1^1d_3^2 + d_3^1d_1^2 & d_2^1d_3^2 + d_3^1d_2^2 & d_3^1d_3^2 & 0 & 0 \\ 2d_1^2d_3^2 & 2d_2^2d_3^2 & (d_3^2)^2 & 0 & -1 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \\ \zeta \\ (\sigma')^2 \\ (\sigma'')^2 \end{pmatrix} = - \begin{pmatrix} (c_1^1)^2 + (c_2^1)^2 \\ c_1^1c_1^2 + c_2^1c_2^2 \\ (c_1^2)^2 + (c_2^2)^2 \\ (d_1^1)^2 + (d_2^1)^2 \\ d_1^1d_1^2 + d_2^1d_2^2 \\ (d_1^2)^2 + (d_2^2)^2 \end{pmatrix}. \tag{33}$$

Solving this system of equations in the least squares sense gives $(\xi, \eta, \zeta, (\sigma')^2, (\sigma'')^2)$ as function of c_i^j and d_i^k determined from (27) and (28). Then, Γ is given by

$$\Gamma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \xi & \eta & \pm\sqrt{\zeta - \xi^2 - \eta^2} \end{pmatrix}, \tag{34}$$

and we estimate the first two rows of R' in (30) by $\tilde{N}' = \sigma' N'\Gamma$. The third row is then easily obtained as the cross product of the first two rows: ${\rho'}^3 = {\rho'}^1 \times {\rho'}^2$. The ambiguity in the determination of γ_3 in Γ corresponds to a sign change in the last component of the first two rows of R' and R'', and a corresponding sign change in the last row, *i.e.*, the following solutions:

$$\rho = \begin{pmatrix} \rho_1^1 & \rho_2^1 & \rho_3^1 \\ \rho_1^2 & \rho_2^2 & \rho_3^2 \\ \rho_1^3 & \rho_2^3 & \rho_3^3 \end{pmatrix}, \qquad \bar{\rho} = \begin{pmatrix} \rho_1^1 & \rho_2^1 & -\rho_3^1 \\ \rho_1^2 & \rho_2^2 & -\rho_3^2 \\ -\rho_1^3 & -\rho_2^3 & \rho_3^3 \end{pmatrix}. \tag{35}$$

This ambiguity reflects the fact that for scaled orthographic projection we cannot distinguish between a positive rotation of a point in front of the center of rotation and a negative rotation of a similar point behind the center of rotation. To choose between the two possible solutions, we can either assume similarity between adjacent rotations, or use the size variations of the tracked image features (see section 6.2).

The matrices obtained from (35) depend upon Γ and $(\xi, \eta, \zeta, (\sigma')^2, (\sigma'')^2)$ and are not guaranteed to be orthogonal matrices, since $(\xi, \eta, \zeta, (\sigma')^2, (\sigma'')^2)$ is computed from an overdetermined system of equations. Given an estimate ρ of the rotation matrix R, a singular value decomposition is carried out of ρ, and R is determined from $\rho = U\Sigma V^T$, which gives $R = UV^T$. This choice minimizes the difference between ρ and R in the Frobenius norm.

6 Joint factorization of point and line correspondences

The treatment so far shows how changes in the orientation of an unknown three-dimensional point and line configuration can be computed from three affine views. To derive corresponding motion descriptors from time sequences, we shall in this section develop a factorization approach, which treats point features and line features together. In this way, we shall combine several of the ideas in the factorization methods for either point features or line features (Tomasi and Kanade [14], Quan and Kanade [16], Sturm and Triggs [17]). It should be noticed, however, that the main intention here is not to separate the motion information from structural information a priori as in Tomasi and Kanade [14]. The goal is to exploit the redundancy between point features and line features over multiple frames, and to avoid the degenerate cases that are likely to occur if we compute three-dimensional motion using image triplets only.

Let us introduce a slightly different notation (and do away with the Einstein summation convention). The centered affine projection of a three-dimensional point $P_k = (X_k, Y_k, Z_k)^T$ in image n shall be written

$$\begin{pmatrix} x_k^n \\ y_k^n \end{pmatrix} = M^n P_k = \begin{pmatrix} - & \alpha^{nT} & - \\ - & \beta^{nT} & - \end{pmatrix} \begin{pmatrix} X_k \\ Y_k \\ Z_k \end{pmatrix}, \tag{36}$$

while the (centered) affine projection of a line $P_l = (X_{l,0}, Y_{l,0}, Z_{0,l})^T + \tau (U_l, V_l, W_l)^T = P_{l,0} + \tau Q_l$ in image n shall be represented by the directional vector

$$\lambda_l^n \begin{pmatrix} u_l^n \\ v_k^n \end{pmatrix} = M^n Q_l = \begin{pmatrix} - & \alpha^{nT} & - \\ - & \beta^{nT} & - \end{pmatrix} \begin{pmatrix} U_l \\ V_l \\ W_l \end{pmatrix}, \tag{37}$$

where the suppression of $(X_{l,0}, Y_{l,0}, Z_{0,l})^T$ and the introduction of the scale factor λ_l^n account for the fact that the position of the line is unimportant, the length of (u_l^n, v_k^n) is unknown, and only the orientation of the line is significant. Given K point and L line correspondences over N image frames, we model these measurements together by a matrix $G = MS$ according to

$$\begin{pmatrix} x_1^1 & \dots & x_K^1 & \lambda_1^1 u_1^1 & \dots & \lambda_L^1 u_L^1 \\ y_1^1 & \dots & y_K^1 & \lambda_1^1 v_1^1 & \dots & \lambda_L^1 v_L^1 \\ \vdots & & \vdots & \vdots & & \vdots \\ x_1^N & \dots & x_K^N & \lambda_1^N u_1^N & \dots & \lambda_L^N u_L^N \\ y_1^N & \dots & y_K^N & \lambda_1^N v_1^N & \dots & \lambda_L^N v_L^N \end{pmatrix}$$
$$= \begin{pmatrix} - & \alpha^{1T} & - \\ - & \beta^{1T} & - \\ & \vdots & \\ - & \alpha^{NT} & - \\ - & \beta^{NT} & - \end{pmatrix} \begin{pmatrix} X_1 & \dots & X_K & U_1 & \dots & U_L \\ Y_1 & \dots & Y_K & V_1 & \dots & V_L \\ Z_1 & \dots & Z_K & W_1 & \dots & W_L \end{pmatrix}. \tag{38}$$

Since the rank of the matrices on the right hand side is maximally three, it follows that any 4×4-minor must be zero, and we can, for example, form selections of $k, k', k'' \in$

$[1..K]$, $l \in [1..L]$ and $n, n' \in [1..N]$, with

$$\begin{vmatrix} x_k^n & x_{k'}^n & x_{k''}^n & \lambda_l^n u_l^n \\ y_k^n & y_{k'}^n & y_{k''}^n & \lambda_l^n v_l^n \\ x_k^{n'} & x_{k'}^{n'} & x_{k''}^{n'} & \lambda_l^{n'} u_l^{n'} \\ y_k^{n'} & y_{k'}^{n'} & x_{k''}^{n'} & \lambda_l^{n'} v_l^{n'} \end{vmatrix} = 0. \tag{39}$$

If we would have $K \geq 4$ point correspondences, this would give us up to $\binom{K}{3}\binom{N}{2}L$ linear relationships, out of which a subset could be selected for determining the scale factors λ_l^n from an overdetermined system of homogeneous linear equations. Approaches closely related to this have been applied to line features by Quan and Kanade [16] and to point features by Sturm and Triggs [17].

Given only three points, however, these linear relationships degenerate, since any minor with $K = 3$ point features is zero (due to centering, all the K points together will be linearly dependent). The same thing happens when the points are coplanar, see section 7.

To determine λ_l^n (totally NL scaling factors) in this case, we instead apply the affine trifocal tensor to a set of randomly selected triplets of image frames as a preprocessing stage. In analogy with Quan and Kanade [16] let us for each such triplet $n, n', n'' \in [1..N]$, insert the following shape matrix

$$\begin{pmatrix} X_1 & X_2 & X_3 \\ Y_1 & Y_2 & Y_3 \\ Z_1 & Z_2 & Z_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{40}$$

into the projection equation (38) for $K = 3$ point features:

$$H^{n,n',n''} = \begin{pmatrix} - & \alpha^{nT} & - & \lambda_1^n u_1^n & \dots & \lambda_L^n u_L^n \\ - & \beta^{nT} & - & \lambda_1^n v_1^n & \dots & \lambda_L^n v_L^n \\ - & \alpha^{n'T} & - & \lambda_1^{n'} u_1^{n'} & \dots & \lambda_L^{n'} u_L^{n'} \\ - & \beta^{n'T} & - & \lambda_1^{n'} v_1^{n'} & \dots & \lambda_L^{n} v_L^{n'} \\ - & \alpha^{n''T} & - & \lambda_1^{n''} u_1^{n''} & \dots & \lambda_L^{n''} u_L^{n''} \\ - & \beta^{n''T} & - & \lambda_1^{n''} v_1^{n''} & \dots & \lambda_L^{n} v_L^{n''} \end{pmatrix}. \tag{41}$$

Then, since the rank of the right hand side in (38) is maximally three, it follows that any 4×4-minor of this matrix must be zero. For each line feature $l \in [1..L]$, we consider three algebraically independent minors. Given three camera matrices M^n, $M^{n'}$ and $M^{n''}$, these minors define three homogeneous linear relations between λ_l^n, $\lambda_l^{n'}$ and $\lambda_l^{n''}$ for each $l \in [1..L]$. The camera matrices for stating these relations are determined by computing the trifocal tensor for the corresponding triplets of image features as described in section 5.

From a set of such (randomly selected) triplets, we then for each l define a homogeneous system of equations of the following type for determining λ_l^n:

$$D_l \Lambda_l = \begin{pmatrix} * & \dots & * \\ \vdots & \ddots & \vdots \\ * & \dots & * \end{pmatrix} \begin{pmatrix} \lambda_l^1 \\ \vdots \\ \lambda_l^N \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{42}$$

Three consecutive rows in D_l correspond to one image triplet, and the entries in the matrix D_l have just been indicated by '$*$' symbols. In practice, we let the number of triplets be substantially larger than the number of image frames (by a factor 4). Moreover, the image triplets are ranked by sorting and thresholding with respect to a condition number.

Then, Λ_l is determined from the overconstrained system of equations using a singular value decomposition of $D_l = U_l \Sigma_l V_l^T$, which gives Λ_l as the last row of V_l. The λ_l^n values are inserted into G in (38) and a singular value decomposition is computed $G = U_G \Sigma_G V_G^T$. With s_{Gi} denoting the singular values of s_G, all elements except the three first ones in s_G are set to zero to reduce the rank to three. In other words, $\tilde{s}_G = \mathrm{diag}(s_1, s_2, s_3, 0, \ldots, 0)$ gives $\tilde{G} = U_G \tilde{\Sigma}_G V_G^T$. Finally, the ambiguity in the separation of motion information from structure information $G = MS = \hat{M}LL^{-1}\hat{S}$ is resolved in a similar fashion as in (Tomasi and Kanade [14], Quan and Kanade [16]). In this way, refined estimates are obtained for the rotation matrices of the motion as well as the structure of the object.

6.1 Structure estimation from point and line correspondences

The ambiguity $\tilde{G} = MS = \hat{M}LL^{-1}\hat{S}$ in the separation of the motion information from the structure information in $\tilde{G}$ is resolved by forming the matrix $MM^T = (\hat{M}L)(\hat{M}L)^T = \hat{M}LL^T\hat{M}$, which according to (30) is of the following form

$$MM^T = \hat{M}LL^T\hat{M} = \begin{pmatrix} \sigma_1^2 & 0 & & & & \\ 0 & \sigma_1^2 & & & & \\ & & \ddots & & & \\ & & & \ddots & & \\ & & & & \sigma_N^2 & 0 \\ & & & & 0 & \sigma_N^2 \end{pmatrix}. \tag{43}$$

With

$$M^n = \begin{pmatrix} - & \alpha^{nT} & - \\ - & \beta^{nT} & - \end{pmatrix}, \quad \hat{M}^n = \begin{pmatrix} - & \hat{\alpha}^{nT} & - \\ - & \hat{\beta}^{nT} & - \end{pmatrix} \quad \text{and} \quad \hat{M}^n L = M^n \tag{44}$$

where $\alpha^{nT}\alpha^n = \beta^{nT}\beta^n = \sigma_n^2$ and $\alpha^{nT}\beta^n = 0$, we obtain the following system of equations

$$\begin{cases} \hat{\alpha}^{nT}LL^T\hat{\alpha}^n - \hat{\beta}^{nT}LL^T\hat{\beta}^n = 0, \\ \hat{\alpha}^{nT}LL^T\hat{\beta}^n = 0. \end{cases} \tag{45}$$

After introducing

$$A = LL^T = \begin{pmatrix} a_1 & a_2 & a_3 \\ a_2 & a_4 & a_5 \\ a_3 & a_5 & a_6 \end{pmatrix} \quad \text{and} \quad a = (a_1, a_2, a_3, a_4, a_5, a_6)^T \tag{46}$$

let us write (45) as $Ba = 0$, where the components in B are obtained from $\hat{\alpha}^n$ and $\hat{\beta}^n$ in $\hat{M}^n$. This system of equations is solved by singular value decomposition of $B = U_B \Sigma_B V_B^T$, and a is given by the last row in V_B. The matrix A is formed from the vector a, and assuming that this matrix is positive definite, it is diagonalized

$$A = C\Sigma_A C^T = (C\Sigma_A^{1/2})(C\Sigma_A^{1/2})^T \tag{47}$$

to give $L = C\Sigma_A^{1/2}$. This gives the camera matrix $M = \hat{M}(C\Sigma_A^{1/2})$ and the structure matrix $S = (C\Sigma_A^{1/2})^{-1}\hat{S}$. Since M is known, the scaling factors can be determined.

This procedure is a generalization of the factorization method in Tomasi and Kanade [14] from orthographic projection to scaled orthographic projection, see also Poelman and Kanade [15].

6.2 Resolving the ambiguity in the rotation estimates

As described in section 5, the affine projection model gives two possible solutions when determining the rotation R according to (35). A corresponding ambiguity exists for the structure S. These alternatives to the rotation R and the structure S correspond to a mirroring in the z-plane and are denoted by $\bar{R}$ and $\bar{S}$ respectively.

This ambiguity is resolved as follows, using scale information from the scale-space extrema in the multi-scale feature detection step [38] and the scaling factors σ associated with the camera matrices (according to the computational procedure section 6.1). One choice of a rotation matrix R gives a structure for the entire point configuration according to section 6.1. Column number k in S, denoted S_k, gives the structure s_k for point number k. With R' and R'' denoting the rotations at two different moments, the depths z'_k and z''_k of point k are given by

$$\begin{pmatrix} x'_k \\ y'_k \\ z'_k \end{pmatrix} = R' s_k, \qquad \begin{pmatrix} x''_k \\ y''_k \\ z''_k \end{pmatrix} = R'' s_k. \tag{48}$$

Let t'_k and t''_k denote the scales of the corresponding point features, and let the scaling factors of the camera matrices computed according to section 6.1 be σ' and σ''. If the depth z'_k is greater that z''_k, then the relative increase in scale between these two images must be greater than the relative increase in scale of the entire configuration. In other words, one of the conditions

$$\left((z''_k > z'_k) \text{ and } \left(\frac{t''_k}{t'_k} > \frac{\sigma''}{\sigma'}\right)\right) \quad \text{or} \quad \left((z''_k < z'_k) \text{ and } \left(\frac{t''_k}{t'_k} < \frac{\sigma''}{\sigma'}\right)\right) \tag{49}$$

must be satisfied if the structure S and the rotations R' and R'' are correct. Otherwise, we choose the other solution ($\bar{S}$, $\bar{R}'$, $\bar{R}''$) corresponding to a simultaneous mirroring of the structure and all rotations. These conditions are tested for each point in the configuration, and a voting over all points determines which rotation (and thereby structure) should be selected.

6.3 Relative weighting of point and line constraints

In the scheme for structure and motion estimation, we solve overdetermined systems of equations (i) when computing the affine trifocal tensor from the point and line constraints (25) and (26), (ii) when determining the scale factors of the lines from (42) and (iii) in the joint factorization of point and line measurements into structure and motion information (38). When solving such overdetermined systems of equations, it is of crucial importance that the equations are properly weighted. For example, the SVD solution of a homogeneous system of equations is a maximum-likelihood-solution if the noise in the components is independent and has an isotropic Gaussian distribution.

In our case, the equations we solve involve point and line coordinates of automatically extracted image features. Unfortunately, we cannot assume the noise terms in the coordinates and the relations to be independent, since they originate from highly coupled measurements in the image domain. Our aim will therefore be to renormalize the equations by weighting the components to have approximately the same variance, see Shapiro [44], Kanatani [45] and Mühlich and Mester [46] for related works with similar aims.

As prerequisite for this analysis, we can observe that the (centered) 2D point coordinates (x, y) will be in the range $[-\Delta, \Delta]$. Concerning the line coordinates (l_1, l_2, l_3), we have chosen to normalize the directional vector (l_1, l_2) such that $\sqrt{l_1^2 + l_2^2} = 1$. Then, $|l_3|$ equals the orthogonal distance from the line to the origin and will be in the range $[0, \sqrt{2} * \Delta]$. The assumptions we make about the data are that the variance of the errors in x, y and l_3 are of the same order of magnitude and that the errors in l_1 and l_2 have equal variance.

6.3.1 Computing the centered affine trifocal tensor

When computing the elements of the centered affine trifocal tensor from the four point constraints (25) and the three line constraints (26), we solve a system of the form $Ax = 0$. In the first two line constraints, the l_3 components occurs once in each term, while in the last constraint the l_3 component does not occur at all. Motivated by this risk for underweighting of the third line relation, and the fact that the four point relations together contain 16 terms, while the third line relation contains 8 terms, we have chosen to reweight the third line relation such that

$$\frac{||A_p||_{Fro}}{\sqrt{16P}} = \frac{||A_{l3}||_{Fro}}{\sqrt{8L}}, \tag{50}$$

where A_p are the rows of A corresponding to the point relations, A_{l3} are the rows corresponding to the third line relation, and P and L represent the numbers of points and lines.

To demonstrate the importance of this reweighting, table 1 shows a synthetic experiment with 3 equidistant points and 3 lines, distributed such that all features were not in the same plane, while all the lines passed through the centroid of the 3 points. From a random initial pose, this configuration was subjected to a rotation around the vertical Y-axis. three orthographic projections of the configuration were selected, and the image features were perturbed in a similar way as previously. The rotation was estimated using the centered affine trifocal tensor and measures of the errors in the rotation angles were computed as averages over 10 experiments. In this case, the reweighting decreases the error by one order of magnitude. A more thorough normalization could involve a reweighting of point and line constraints, based on covariance estimates of the image features.

6.3.2 Finding scale factors of lines prior to factorization

When computing the scale factors Λ_l of the lines from (42), we can observe that the left hand side matrix is sparse and some columns have few entries due to the random selection of image triplets. The system is solved using SVD, finding the eigenvector of $D_l^T D_l$ that corresponds to the smallest eigenvalue. Noisy data combined with sparse entries of D_l could cause $\min ||D_l \Lambda_l||_2$ to approximately equal the length of the

Rotation error (degrees)	$\Delta\theta$	$\Delta\phi$
Weighted	2.03	2.83
Unweighted	35.2	17.4

Table 1: Synthetic experiment with 3 points and 3 lines at noise level $\nu = 0.01$, showing the accuracy in rotation estimates with and without the reweighting of the third line equation. The results are averages over 10 experiments.

shortest column vector of D_l, thus enforcing a solution vector Λ_l with one dominant element. We avoid this selection of the shortest column vector by normalizing all column vectors to have unit length prior to the SVD computation and afterwards multiply the elements of the solution vector by the normalization factors. This treatment clearly violates the assumptions of equal variance, but produces more stable solutions.

6.3.3 Simultaneous factorization of points and lines

In the simultaneous factorization of points and lines, it is important to treat the point and line data approximately equally. In order to obtain approximately the same error distribution in all columns of the measurement matrix G in (38), we divide the columns corresponding to the point features in G by a factor w,

$$w = \frac{\sqrt{L} * ||G_p||_{Fro}}{\sqrt{P} * ||G_L||_{Fro}} \tag{51}$$

where G_P is the part of G holding the point feature coordinates, G_L is the part holding the direction of the line features and P and L are the number of points and lines. After factorization into $G = MS$, we multiply the part of the structure matrix S that holds the 3D positions of the point features by w. This gives the same l_2 norm of all columns.

To show the importance of this column weighting, we made synthetic experiments where four random points and lines were rotated around the vertical Y-axis. Orthographic projections of these configuration (of size about 200 pixels) were computed with 4 degrees of rotation between each frame, and the image features were perturbed as before. The rotation was estimated using the proposed factorization scheme and the errors in the rotation after 30 frames are shown in table 2 as averages over 20 experiments. (Here, the scale factors of the lines were determined directly after 30 frames, and not in the iterative way that will be described later in section 8.1.) As can be seen, the errors in this case decrease by a factor of 5.

Rotation error (degrees)	$\Delta\theta$	$\Delta\phi$
Weighted	1.24	2.24
Unweighted	6.89	6.73

Table 2: Synthetic experiment with 4 points and 4 lines at noise level $\nu = 0.02$, showing the rotation errors with and without the proposed relative weighting of point and line columns. The results are averages over 20 experiments.

7 Degenerate situations

So far, we have assumed that all point and line configurations are generic. More generally, one may ask what are the degenerate situations, *i.e.*, the cases where not all image features provide additional information to the structure and motion estimation. The subject of this section is to analyse such degeneracies. The methodology we shall follow is to consider the matrix G with measured point and line features according to (38), factorized as the product $G = MS$ of a matrix M with motion parameters and a shape matrix S. Since the rank of any of these matrices is maximally three, we can treat 3×3-minors of lower rank as degenerate configurations.

7.1 Degenerate three-dimensional shapes

With respect to the structure matrix, we can thus distinguish four different cases, depending on the number of point and line features, respectively, we include when forming 3×3-minors:

Three points A minor with three points in three dimensions

$$\begin{vmatrix} X_1 & X_2 & X_3 \\ Y_1 & Y_2 & Y_3 \\ Z_1 & Z_2 & Z_3 \end{vmatrix} \tag{52}$$

is degenerate if the plane through these three points contains the origin. Special cases of this condition include when (i) one of the points is at the origin, (ii) all three points are on the same line, (iii) two of the points are on a line through the origin, or (iv) two or more points coincide.

The rationale for the special treatment of the origin is that under affine projection the motion of the center of gravity of a three-dimensional point configuration is given by the motion of the center of gravity of the two-dimensional image measurements. Using centered coordinates, the center of gravity coincides with the origin, and we can regard the origin as one measurement implicitly present with all the other measurements.

Two points and one line A minor with two points and one line

$$\begin{vmatrix} X_1 & X_2 & U_3 \\ Y_1 & Y_2 & V_3 \\ Z_1 & Z_2 & W_3 \end{vmatrix} \tag{53}$$

is degenerate if the line is contained in the plane through the two points and the origin. Special cases of this condition (excluding previously listed degeneracies) include when: (i) the line is parallel to the line through the two points, and (ii) the line is parallel to the line through the origin and one of the points.

One point and two lines A minor with one point and two lines

$$\begin{vmatrix} X_1 & U_2 & U_3 \\ Y_1 & V_2 & V_3 \\ Z_1 & W_2 & W_3 \end{vmatrix} \tag{54}$$

is degenerate when the point is contained in the plane spanned by the two lines and the origin. One special case of this is when the two lines are parallel.

Three lines A minor with three lines

$$\begin{vmatrix} U_1 & U_2 & U_3 \\ V_1 & V_2 & V_3 \\ W_1 & W_2 & W_3 \end{vmatrix} \tag{55}$$

is degenerate if the three lines are all in the same plane. For this situation, one could at first expect analogous degeneracies as for the abovementioned situation with three lines. The situation is different for lines, however, for the following reasons: (i) the length of each line is non-zero, (ii) the end points of three lines cannot be on the same line if we assume that the lines are normalized to unit length, (iii) the end points of two normalized lines are on a line through the origin only if they correspond to opposite directions.

Remarks From an intuitive viewpoint, one would also expect that situations where a line that goes through a feature point or two lines that intersect at a feature point would be regarded as degenerate cases. These degeneracies are, however, not covered by this analysis, since the affine projection model in (38) does not take the position of the line into account. On one hand, the change in orientation of a line under an affine deformation is independent of the position of the line, motivating the orientation parameterization of the line features (37). On the other hand, lines do not reposition themselves randomly when subject to affine deformations. The latter effect is not explicitly modelled in the present factorization method, while the positions of the lines are included (as the third coordinate of the line coordinates) in the point relations arising from the affine trifocal tensor (26).

From this analysis, we can conclude that point and line features extracted from man-made objects will often lead to a high level of degeneracy concerning the mutual relations between point and line features. Hence, image measurements may not contribute as much *geometrically* to the problem as one might expect from a generic viewpoint. Nevertheless, such measurements can be expected to contribute *statistically*, in a similar manner as multiple measurements of the same physical structure may reduce the effective noise level in the image measurements.

7.2 Degenerate three-dimensional motions

To determine degenerate motions, let us study 3×3 minors of the matrix M in (38), keeping in mind that for scaled orthographic projection the camera matrix corresponds to a rescaling of the first two rows of a rotation matrix.

For a pure rotation around the optical axis (the Z-axis), the third column of the camera matrix will always be zero. This means that all the minors of M will be zero, the assumption of G having rank three is violated, and we cannot recover the 3D structure of the object.

8 Experiments

When using the abovementioned methodology for estimating the structure and motion of an unknown object, we can expect the convergence properties and the numerical accuracy to be influenced by several factors: (i) the three-dimensional structure of the object, (ii) its three-dimensional motion, (iii) the validity of the affine approximation of the perspective mapping, and (iv) the localization errors of the image features. Specifically, we can expect that the numerical accuracy will increase with the number of available image features as well as how many independent views are seen. Moreover, we can expect the performance to decrease as the perspective effects become larger and if the localization errors are large compared to the interframe motions. Ideally, one would like to have compact closed-form expressions that characterize how the different types of errors propagate from the input to the output. Since, however, we can expect such a theoretical analysis to be rather complex, we will in this section present a systematic experimental study, to determine empirical performance bounds for each one of the abovementioned factors.

8.1 Experiments on synthetic test data

To investigate the properties of the abovementioned framework for computing structure and motion, we shall first carry out investigations on synthetic data, which are generated by the following procedure: A three-dimensional point and line model is generated. Two types of synthetic test objects will be considered: (i) random selection of K points and L lines from a Gaussian distribution, and (ii) a qualitative hand model with four fingers. The intention behind the first choice is to consider a large number of different shapes, such that the performance values will not be specifically shape dependent. The second test object is selected because of the specific application in vision-based human-computer interaction we are interested in.

The experimental protocol we will follow is to subject each test object to a smooth three-dimensional rotation around a fixed axis. For each frame, a perspective projection is computed, and noise is added in the image domain. For point features, the positions of the image features are perturbed by additive white Gaussian noise with standard deviation Σ, determined to be proportional to the size of the object in the image domain, measured as a factor ν times the maximum distance between the point features to the centroid of all the image points. For line features, the disturbances are introduced by representing each line by two endpoints, and then disturbing the two end points independently. All lines have the same length in three dimensions, and the uncertainty will therefore be relatively higher for lines that are parallel to the viewing direction.

Concerning the computation of the scale factors of the lines we will, unless otherwise stated, use the tensor-based method described earlier. For each consecutive frame, the set of equations is increased iteratively, by adding new equations derived from new triplets of images, where each new triplet includes the present frame. In this way, we can reuse equations and avoid the computation of a completely new set of equations for each frame.

8.1.1 Error measures

Motion estimates To quantify the errors in the estimated rotations, we measure the angle $\Delta\theta$ between the real eigenvector u of the true rotation matrix and the real eigenvector v of the estimated rotation matrix, as well as the difference $\Delta\phi$ between the estimated and the real rotation around this rotation axis. In graphs, we usually display the Euclidean sum of θ and ϕ, $\Delta\varphi = \sqrt{\Delta\theta^2 + \Delta\phi^2}$, which serves as an upper bound on either $\Delta\theta$ or $\Delta\phi$.

Structure estimates For points, we quantify errors in the structure estimates by the Euclidean sum of the difference between the true and the estimated three-dimensional points coordinates $(X_k, Y_k, Z_k)^T$, respectively,

$$d_P(\hat{X}, X) = \left\| \begin{matrix} \hat{X}_1 - X_1 & \dots & \hat{X}_K - X_K \\ \hat{Y}_1 - Y_1 & \dots & \hat{Y}_K - Y_K \\ \hat{Z}_1 - Z_1 & \dots & \hat{Z}_K - Z_K \end{matrix} \right\|_2 , \tag{56}$$

and for lines by the Euclidean sum of the difference between the true and the estimated normalized line coordinates $(U_l, V_l, W_l)^T$ with $U_l^2 + V_l^2 + W_l^2 = 1$,

$$d_L(\hat{U}, U) = \left\| \begin{matrix} \hat{U}_1 - U_1 & \dots & \hat{U}_L - U_L \\ \hat{V}_1 - V_1 & \dots & \hat{V}_L - V_L \\ \hat{W}_1 - W_1 & \dots & \hat{W}_L - W_L \end{matrix} \right\|_2 . \tag{57}$$

To make the structure errors invariant to scalings and rotations, the point structure matrices are first normalized to unit Frobenius norm. Then, the estimated structure is aligned to the true structure using the rotation that minimizes the above point structure error measure. This normalization corresponds to introducing the following relative error measures

$$\epsilon_{SP} = \frac{d_P(\hat{X}, X)}{d_P(\hat{X}, 0)}, \qquad \epsilon_{SL} = \frac{d_L(\hat{U}, U)}{d_L(\hat{U}, 0)}, \tag{58}$$

which are less specifically dependent on the size of the object and the number of object features than $d_P(\hat{X}, X)$ and $d_L(\hat{U}, U)$. In graphs, we shall often display the composed structure error measure $\epsilon_{SC} = \epsilon_{SP} + \epsilon_{SL}$.

8.1.2 Influence of feature localization errors

To investigate the influence of noise on the convergence properties and the accuracy of the structure and motion estimates, we first consider synthetic data generated by an orthographic projection model. Four points and four lines were randomly selected from a Gaussian distribution, this object was subject to a rotation around the vertical Y-axis with a rotation of 4 degrees between successive frames. For each frame, an orthographic projection was computed and noise was added to the image features, with standard deviation proportional to the size of the object as described above. Three different noise levels $\nu = 0.02$, 0.05 and 0.10 were investigated, basically corresponding to localization errors with standard deviations of 2 pixels, 5 pixels and 10 pixels, respectively, if we assume that the object occupies 200 pixels in the image domain. For a smaller size object occupying say 50 pixels, these noise levels correspond to localization errors of about half a pixel, one pixel and two pixels, respectively.

For each noise level, this procedure was repeated for 10 randomly selected configurations, and figure 2 shows the average rotation error measure $\Delta\varphi$, the average point structure error measure ϵ_{SP} and the average line structure error measure ϵ_{SL} at each frame. As can be seen, the error measures decrease rapidly with the number of image frames, reflecting the fact that the motion and structure estimates become more accurate as more views of the object have been seen. (To avoid the initial transient effects, the calculations were started only after 10 frames, when the object had rotated totally 40 degrees.) Specifically, the rotation error measure decreases in a similar way as the point structure measure. In view of the fact that the structure of the object has lower degrees of freedom than all its rotation states, we can thus interpret an accurate estimation of the object shape as a prerequisite for computing accurate object pose. Moreover, the error measures reach an approximate steady-state after about 25-30 frames, when the object has rotated by altogether 90-120 degrees. The error measures in steady-state are roughly proportional to ν.

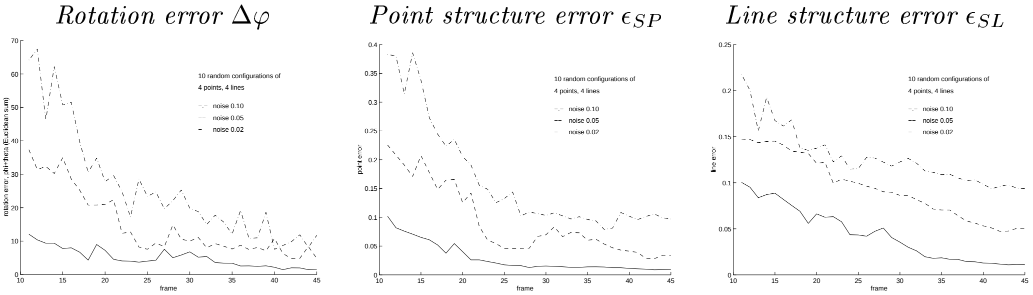

Rotation error $\Delta\varphi$ *Point structure error* ϵ_{SP} *Line structure error* ϵ_{SL}

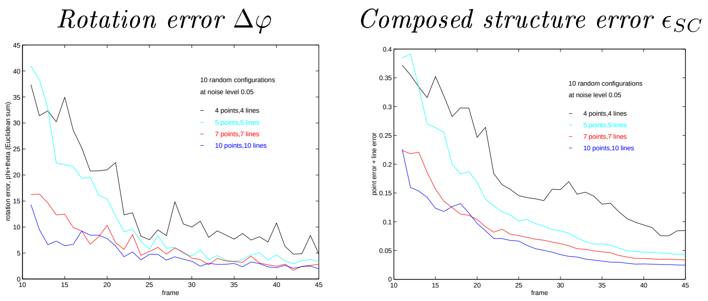
Figure 2: Influence of noise in the image domain on rotation and structure estimates for the orthographic projection of synthetic test objects with four points and lines. The error measures are averages over 10 random selections of points and lines from a thresholded normal distribution. The rotation is 4 degrees per frame. Note how the accuracy increases with the object motion, and see the text for further explanations.

8.1.3 Influence of number of image features

To investigate the influence of the number of image features, we then varied the number of image features and selected 4, 5, 7 and 10 random points and lines, respectively, from a Gaussian distribution. In all other respects, the experimental conditions were the same as in section 8.1.2. Figure 3 shows the results for the noise level $\nu = 0.05$. As can be seen, the error measures decrease with the number of image features, indicating that both the rotation and the structure estimates will be more accurate as more image features are available and the overdeterminacy in the equations thus increases. Specifically, the ability of the scheme to converge for highly noisy image data is also higher when the number of image features is large.

8.1.4 Influence of perspective effects

To investigate the influence of perspective effects, let us next replace the orthographic projection model by perspective projection. Initially, we consider a test object with

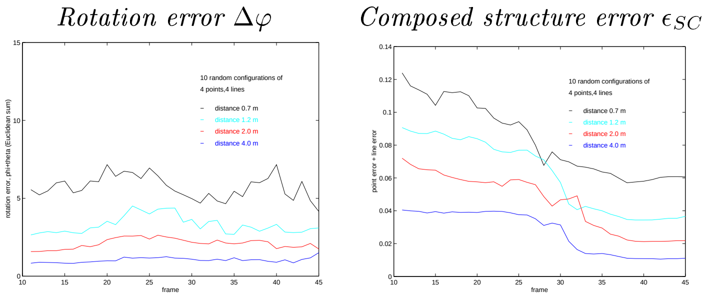

Rotation error $\Delta\varphi$

Composed structure error ϵ_{SC}

Figure 3: Influence of the number of image features on rotation and structure estimates for the orthographic projection of synthetic test objects at noise level $\nu = 0.05$.

4 points and 4 lines having a diameter rescaled to $d = 10$ cm, and viewed at distances of $D = 0.7$ m, 1.2 m, 2.0 m and 4.0 m, respectively. (The focal length of the camera is not important here, since the noise level is proportional to the size of the object in the image.). The intention behind these viewing conditions is to consider an object with the approximate size of a hand viewed by a computer equipped with a camera in an office environment. In all other respects, the experimental conditions are the same as described in section 8.1.3.

Figure 4 shows results in the noise free case, showing how the errors in the rotation and the structure estimates depend on the viewing distance D. Figure 5 shows corresponding results at a fixed viewing distance of $D = 1.2$, while the noise level assumes the values $\nu = 0.0, 0.02, 0.05$ and 0.10. As can be seen, the influence by perspective effects is significant at small viewing distances, while it decreases as the distance to the object gets larger. When we add noise to the input, the errors are initially larger, while in steady-state the scheme reaches a level that roughly corresponds to the maximum of the errors due to noise and perspective effects.


Rotation error $\Delta\varphi$

Composed structure error ϵ_{SC}

Figure 4: Influence of perspective effects on rotation and structure estimates for synthetic test objects without noise. The distance to the object is 0.7, 1.2, 2.0 and 4.0 m.

8.1.5 Influence of temporal sampling density

So far, the three-dimensional motion has been the same in all experiments — a rotation of 4 degrees per frame around the Y-axis. To investigate the influence of the temporal sampling density, we shall in this section simulate a change of temporal sampling by changing the interframe motion to 2, 3 and 8 degrees. The results in

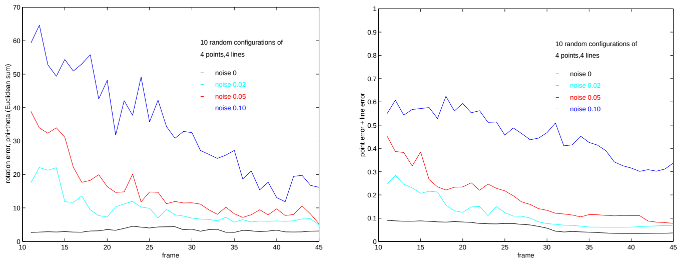
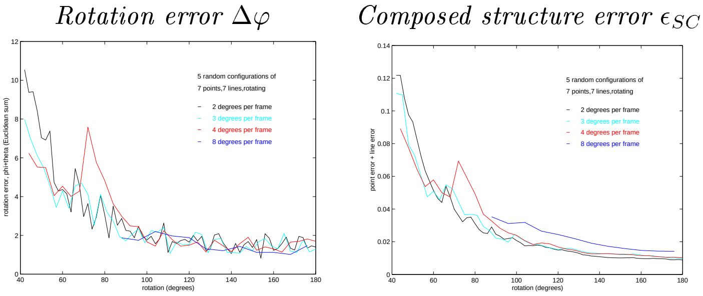


Figure 5: Influence of perspective effects on rotation and structure estimates for synthetic test objects at different noise levels. The distance to the object is 1.2 m, and the number of image features is 4 points and 4 lines.

figure 6 show that besides initial transient effects when the number of image frames is small, the accuracy of the results is mainly determined by the total motion. Hence, we can without serious loss of information expect to be able to use a rather coarse temporal sampling to speed up the computations, once we have an accurate estimate of the object shape. The convergence is, however, slightly faster when the temporal sampling is dense, since more measurements are used in the least squares estimation.


Rotation error $\Delta\varphi$

Composed structure error ϵ_{SC}

Figure 6: Errors in rotation and structure estimates when increasing and decreasing the inter-frame motion, using a test object with 7 points and 7 lines under orthographic projection and noise level $\nu = 0.02$.

8.2 Dependency on object shape

In the previous experiments, we randomly selected points and lines from a (thresholded) normal distribution. By defining test objects from a random process, we ensure that, with probability one, no degenerate situations occur, and that we cover objects of a variety of different shapes. Some configurations will, however, be close to degenerate.

As mentioned in section 7, the general algorithm fails when all the points are in the same plane and all the lines are parallel to this plane. With application to the 3-D hand mouse mentioned in sections 1–2, we have a special interest in investigating under what circumstances the proposed method can estimate the rotation of a human hand. When we hold a hand in a general position, it will often be the case that all the finger tips are in approximately the same plane. (Try this, by putting your finger tips against a table!) Thus, the point configuration will be degenerate, and we cannot use the fast method for computing the scale factors of the lines.

In a first experiment, we want to compare the performance of the fast method (which assumes that no triplet of 3-D points goes through center of gravity of the point configuration) with the more general method based on the affine trifocal tensor. We choose an almost planar test object consisting of 4 points and 4 lines as shown in figure 7. The points are first equally distributed on a circle of diameter 1, thus forming a square. Then, the fourth point is moved to height h above the plane through this circle. The 4 lines go from each one of the 4 object points to a common point of intersection, at the same distance from the points as the distance between them. This object is randomly oriented in 3D before it is rotated around the vertical Y-axis.

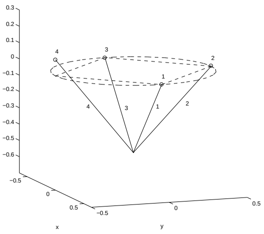
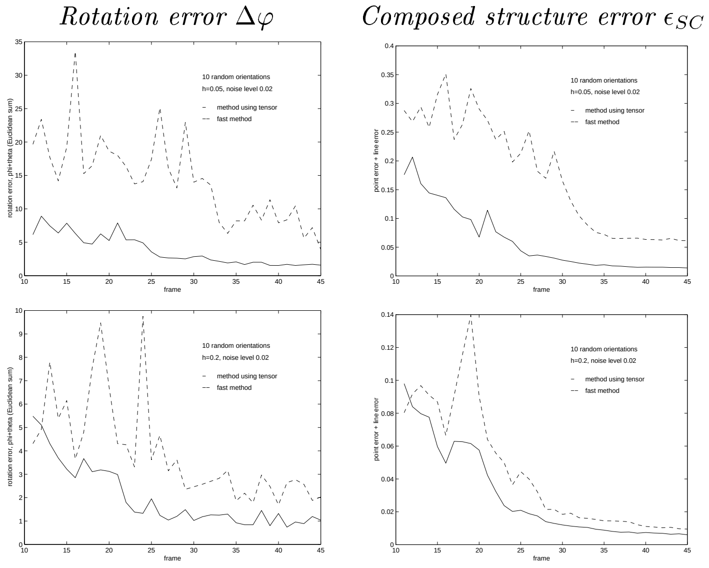


Figure 7: Synthetic test object with a close to planar point-structure.

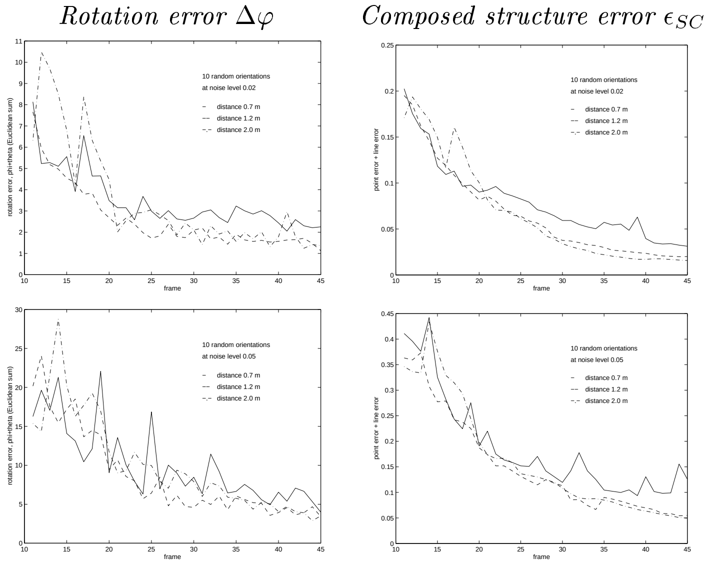
In a first experiment, we let h be 0.05 and 0.2, the noise level is set to 0.02 and the orthographic projection model is used. Figure 8 shows the errors in the estimated structure and rotation when using the two different methods for determining the scale factors of the lines. As we expect, the trifocal tensor-based method is superior when the point-structure is closer to planar ($h = 0.05$), while the difference between the errors of the two methods gets smaller when the point structure gets more non-planar ($h = 0.2$).

In the second experiment, we want to investigate the influence of perspective effects and noise and concentrate on close to planar point structures ($h = 0.05$) using the trifocal tensor-based method for estimating the scale factors of the lines. Figure 9 shows the errors in the rotation and structure estimates for the noise levels $\nu = 0.02$, 0.05 and varying distances 0.7 m, 1.2 m and 2.0 m. At noise level $\nu = 0.02$ the perspective effects are clearly visible, while for $\nu = 0.05$ the effect of the noise dominates over the perspective effects.



Rotation error $\Delta\varphi$ *Composed structure error* ϵ_{SC}

Figure 8: Experiments with the close to planar point structure in figure 7, using two different methods for computing the scale factors of the lines (see the text). The graphs show the errors in rotation and structure estimates for $h = 0.05$ and 0.2. The projection is orthographic and the noise level is $\nu = 0.02$.



Rotation error $\Delta\varphi$ *Composed structure error* ϵ_{SC}

Figure 9: Influence of noise and perspective effects for a synthetic test object with close to planar point structure ($h = 0.05$). The graphs show errors in the rotation and structure estimates for the noise levels $\nu = 0.02$ and 0.05 when the distance to the object is 0.7 m, 1.2 m and 2.0 m.

8.3 Conclusions from the synthetic experiments

The goal of the previous section has been to investigate the influence of set of parameters on the performance of the proposed method for motion estimation from point and line correspondences. From the experiments in figure 6, we can first of all conclude that the amount of rotation is crucial for the convergence, convergence is reached after approximately the same amount of rotation, independently of the inter-frame motion. As expected, the errors in the resulting structure and motion estimates decrease with (i) decreasing noise level, (ii) decreasing perspective effects, (iii) increasing rotation out of the image plane and (iv) increasing number of feature correspondences.

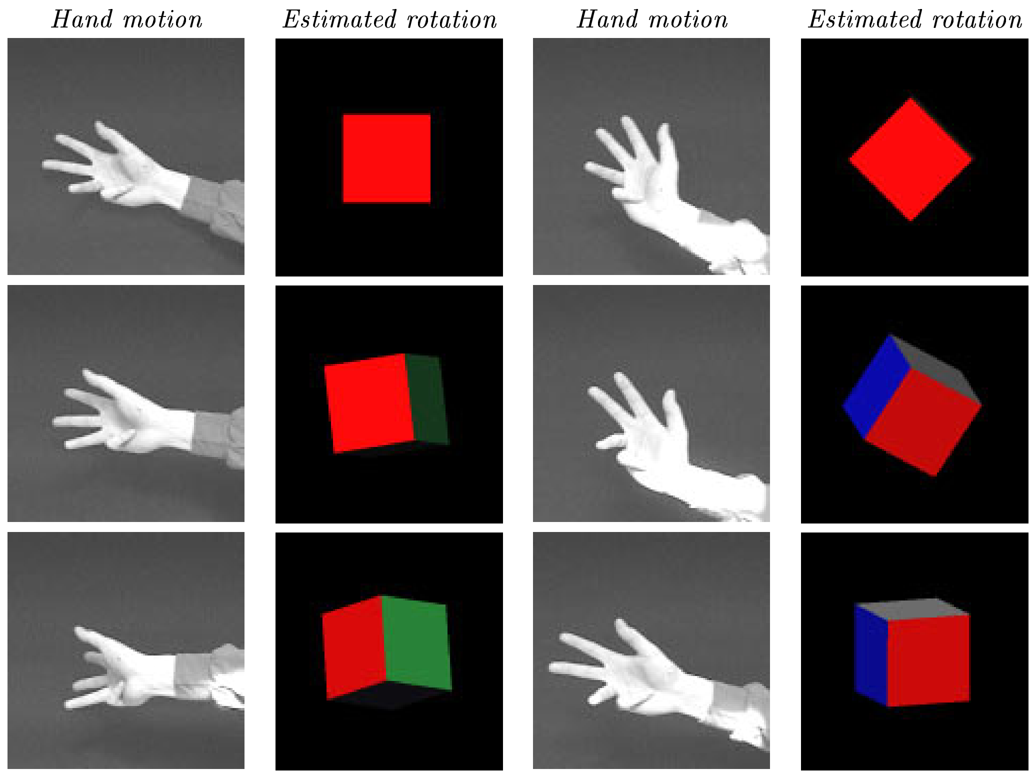
Concerning quantitative numbers, one may ask what conditions must be met for the hand mouse to reach a certain accuracy. By studying the experiments in section 8.1.2 with random configurations, we see that 5 or more points and lines are needed to get rotation errors below 10 degrees when $\nu = 0.10$ or below 5 degrees for $\nu = 0.05$. If we want the rotation error caused by perspective effects to not exceed 5 degrees, the experiments show that the object should be at a distance of more than 10 times the size of the object.

For the specific test object, we can first conclude that the proposed fast method for estimating the scale factors has severe problems. Therefore, the method based on the affine trifocal tensor should be used. Concerning the convergence, we can note that for the noise level $\nu = 0.05$, the rotation around the axis perpendicular to the optical axis has to exceed 90-120 degrees before the rotation error becomes reasonable stable, see figure 9. If the perspective effects are small, and the noise level is less than $\nu = 0.05$ (corresponding to feature localization errors of 5 pixels if the object size in the image is 200 pixels), we can expect the error in the estimated rotation to be below 5 degrees after convergence. A study of the minimal case with 3 points and 3 lines in appendix A.2 shows that we cannot expect a rotation error below 5 degrees if the noise level is above $\nu = 0.02$. This indicates that a minimum of 4 points and 4 lines is preferable for the intended application.

8.4 Experiments on real image data

Our next step is to apply the proposed scheme for structure and motion estimation to feature correspondences computed from real-world image data. As described in section 2, we capture point features corresponding to the finger tips by blob tracking and line features corresponding to the fingers by ridge tracking (see figure 1). Motivated by the results from the synthetic experiments, summarized in section 8.3, we choose to estimate the three-dimensional rotations from the feature trajectories of four fingers and four fingertips.

The left columns in figure 10 show a few snapshots from an image sequence with a hand moving at a distance of 1.0-1.5 m from the camera. Since no ground truth was available in this case, we show the results by subjecting a synthetic cube to the estimated rotations after convergence of the estimated structure. We can see how the motion of the cube mimics the motion of the hand. This effect is more apparent when the images are shown as a temporal sequence.



Hand motion *Estimated rotation* *Hand motion* *Estimated rotation*

Figure 10: 3D rotations estimated from the motion of a human hand. The left column shows the motion of the hand while the right column shows the result of computing changes in 3D orientation using the joint factorization of point and line features detected and tracked in the sequence. The results are illustrated by subjecting a three-dimensional cube to the estimated rotations.

9 Summary and discussion

We have presented a framework for capturing point and line correspondences over multiple affine views. This framework is closely connected to and builds upon several previous works concerning the affine projection model, perspective point correspondences and line correspondences as can be modelled by the trilinear tensor. It also builds upon factorization approaches for affine and perspective projection.

We propose that the (centered) affine trifocal tensor constitutes a canonical tool to model point and line correspondences in triplets of affine views (sections 3–4). This extends the advances of previous works and we show how the trifocal affine tensor relates to the perspective trilinear tensor. Indeed, the algebraic structure of the affine trifocal tensor can be mapped to the algebraic structure of the perspective trilinear tensor. The centered affine trifocal tensor makes it possible to explore sparse sets of point and line features, since it contains 12 non-zero coefficients compared to the 27 coefficients in the trilinear tensor. The computation of motion parameters from the affine trifocal sensor (section 5) is also more straightforward. Closely related formulations of this affine trifocal tensor (Bretzner and Lindeberg [47]) have been presented simultaneously and independently by Kahl and Heyden [48], Quan et al. [49] and Mendonca and Cipolla [50]. More recently, Thórhallsson and Murray [51] have presented an extension to the quadrifocal tensor for modelling affine projections in four affine views, with a follow-up work by Hayman et al. [52].

To capture point and line correspondences in dense time sequences, we have also applied a factorization approach (section 6), to which the affine trifocal tensor serves as an important processing step for computing the scaling factors of line correspondences when three or less point correspondences are available. When four or more point correspondences are given, these scaling factors can be determined directly from a system of linear equations. The performance of this factorization approach has been investigated experimentally (section 8), and its degenerate configurations analysed (section 7). This joint factorization of point and line features was first presented in [47], and a similar approach was developed simultaneously and independently by Kahl and Heyden [48, 53].

The abovementioned theory has been combined with a framework for feature tracking with automatic scale selection (section 2), which has the attractive property that it adapts the scale levels to the local image structure and allows image features to be tracked over large size variations. The extended feature trajectories obtained in this way allow for higher accuracy in the motion estimates, since the relative influence of position errors decreases as the motion gets larger over time. The scale information associated with the image features also resolves the inherent reversal ambiguity of scaled orthographic projection.

Specifically, we have considered a problem in human-computer interaction of transferring three-dimensional orientation to a computer using no other equipment than the operator's own hand (section 2 and section 8.4). Contrary to the more common approach of using detailed geometric hand models, we have here illustrated how changes in three-dimensional orientation can be computed using a qualitative model, based on blob features and ridge features from three fingers. Whereas a more detailed model could possibly allow for higher accuracy in the motion estimates, the simplicity and the generic nature of this module for motion estimation makes it straightforward to implement and lends itself easily to extensions to other problems.

A Appendix

A.1 Algebraic constraints on the affine trifocal tensor

The centered affine trifocal tensor $\mathcal{T}_i^{jk}$ defined in (23) and (24) contains 12 non-zero elements. These elements, in turn, depend on 12 camera parameters c_i^j and d_i^j according to (16) and (17), respectively. The mapping from the camera parameters c_i^j and d_i^j to the affine trifocal tensor $\mathcal{T}_i^{jk}$ is, however, not surjective. This can, for example, be understood from the ambiguity (29) that arises when computing the camera matrices from the affine trifocal tensor. Hence, given a set of 12 arbitrary coeffients, it is not necessarily the case that these coefficients constitute an affine trifocal tensor. In this section, we will derive two algebraic conditions that the elements of the centered affine trifocal tensor must satisfy. Let us insert

$$c_3^1 = -\mathcal{T}_3^{13}, \qquad d_3^1 = \mathcal{T}_3^{31}, \qquad c_3^2 = -\mathcal{T}_3^{23}, \qquad d_3^2 = \mathcal{T}_3^{32}, \tag{59}$$

which we obtain from (24) into the other components of $\mathcal{T}_i^{jk}$ in (24). This gives

$$
\begin{aligned}
T_1^{11} &= c_1^1 T_3^{31} + d_1^1 T_3^{13}, \qquad & T_1^{12} &= c_1^1 T_3^{32} + d_1^2 T_3^{13}, && (60)\\
T_1^{21} &= c_1^2 T_3^{31} + d_1^1 T_3^{23}, \qquad & T_1^{22} &= c_1^2 T_3^{32} + d_1^2 T_3^{23} && (61)\\
T_2^{11} &= c_2^1 T_3^{31} + d_2^1 T_3^{13}, \qquad & T_2^{12} &= c_2^1 T_3^{32} + d_2^2 T_3^{13}, && (62)\\
T_2^{21} &= c_2^2 T_3^{31} + d_2^1 T_3^{23}, \qquad & T_2^{22} &= c_2^2 T_3^{32} + d_2^2 T_3^{23}. && (63)
\end{aligned}
$$

We can eliminate c_i^j from these expressions by forming

$$
\begin{aligned}
T_3^{32}(60a) - T_3^{31}(60b) \quad &\Rightarrow \quad T_1^{11}T_3^{32} - T_1^{12}T_3^{31} = d_1 T_3^{13}T_3^{32} - d_1^2 T_3^{13}T_3^{31}, && (64)\\
T_3^{32}(61a) - T_3^{31}(61b) \quad &\Rightarrow \quad T_1^{21}T_3^{32} - T_1^{22}T_3^{31} = d_1^1 T_3^{23}T_3^{32} - d_1^2 T_3^{23}T_3^{31}, && (65)\\
T_3^{32}(62a) - T_3^{31}(62b) \quad &\Rightarrow \quad T_2^{11}T_3^{32} - T_2^{12}T_3^{31} = d_2^1 T_3^{13}T_3^{32} - d_2^2 T_3^{13}T_3^{31}, && (66)\\
T_3^{32}(63a) - T_3^{31}(63b) \quad &\Rightarrow \quad T_2^{21}T_3^{32} - T_2^{22}T_3^{31} = d_2^1 T_3^{23}T_3^{32} - d_2^2 T_3^{23}T_3^{31}, && (67)
\end{aligned}
$$

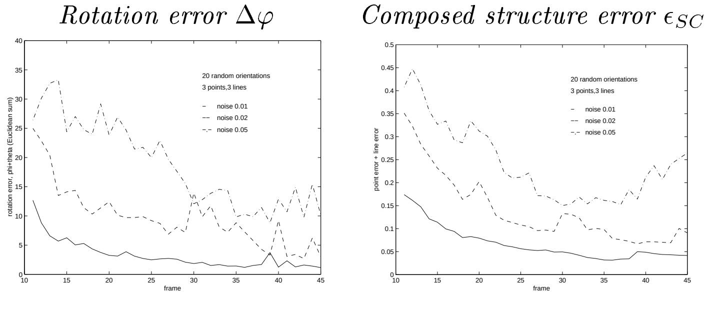
and further eliminate d_i^j by forming

$$
\begin{aligned}
T_3^{23}(64) - T_3^{13}(65) \quad &\Rightarrow \quad T_3^{23}(T_1^{11}T_3^{32} - T_1^{12}T_3^{31}) - T_3^{13}(T_1^{21}T_3^{32} - T_1^{22}T_3^{31}) = 0, && (68)\\
T_3^{23}(66) - T_3^{13}(66) \quad &\Rightarrow \quad T_3^{23}(T_2^{11}T_3^{32} - T_2^{12}T_3^{31}) - T_3^{13}(T_2^{21}T_3^{32} - T_2^{22}T_3^{31}) = 0. && (69)
\end{aligned}
$$

These two relations will be referred to as the trilinear constraints on the centered affine trifocal tensor.

A.2 Experimental investigation of a minimal case

One minimal set of points and line correspondences for computing the affine trifocal tensor consists of 3 points and 3 lines (see section 3.2). Naturally, such a minimal configuration will be very sensitive to degenerate situations and that is the reason why we did not study this case in the experiments in section 8.1 with random configurations of K points and L lines. However, if we choose the first 3 points and the first 3 lines of the synthetic test object in figure 7, we get a minimal configuration with a non-degenerate structure. This configuration was first given a random orientation and was then subjected to a rotation as described in section 8.2. Figure 11 shows the effect of noise on the structure and rotation estimates for orthographic projection


Rotation error $\Delta\varphi$

Composed structure error ϵ_{SC}

Figure 11: Influence of noise in the image domain for the synthetic test object 3 points and 3 lines described in the text. The graphs show errors in the rotation and structure estimates for the noise levels $\nu = 0.01, 0.02$ and 0.05 with orthographic projection.

and noise levels $\nu = 0.01$, 0.02 and 0.05. As can be seen, the errors in the rotation and structure estimates are clearly larger than the corresponding errors for the object with 4 points and 4 lines shown in figure 9, even without perspective effects.

References

[1] S. Ullman, *The Interpretation of Visual Motion.* Cambridge, Massachusetts: MIT Press, 1979.

[2] S. Maybank, *Theory of Reconstruction from Image Motion.* Springer Verlag, New York, 1992.

[3] T. S. Huang and C. H. Lee, "Motion and structure from orthographic projection," *IEEE Trans. Pattern Analysis and Machine Intell.*, vol. 11, no. 5, pp. 536–540, 1989.

[4] T. S. Huang and A. N. Netravali, "Motion and structure from feature correspondences: A review," *Proc. IEEE*, vol. 82, pp. 251–268, 1994.

[5] J. J. Koenderink and A. J. van Doorn, "Affine structure from motion," *J. of the Optical Society of America*, pp. 377–385, 1991.

[6] J. L. Mundy and A. Zisserman, eds., *Geometric Invariance in Computer Vision.* Cambridge, Massachusetts: MIT Press, 1992.

[7] L. S. Shapiro, *Affine analysis of image sequences.* Cambridge, U.K.: Cambridge University Press, 1995.

[8] P. Beardsley, A. Zisserman, and D. Murray, "Navigation using affine structure from motion," in *Proc. 3rd European Conference on Computer Vision*, (Stockholm, Sweden), pp. 85–96, Springer-Verlag, May 1994.

[9] P. McLauchlan, I. Reid, and D. Murray, "Recursive affine structure and motion from image sequences," in *Proc. 3rd European Conference on Computer Vision*, vol. 800, (Stockholm, Sweden), pp. 217–224, Springer-Verlag, May 1994.

[10] P. H. S. Torr, *Motion Segmentation and Outlier Detection.* PhD thesis, Dept. Eng. Science, Univ. of Oxford, U.K., 1995.

[11] O. Faugeras, "Stratification of three-dimensional vision: Projective, affine and metric reconstructions," *J. of the Optical Society of America*, vol. 12, no. 3, pp. 465–484, 1995.

[12] M. E. Spetsakis and J. Aloimonos, "Structure from motion using line correspondences," *Int. J. of Computer Vision*, vol. 4, no. 3, pp. 171–183, 1990.

[13] J. Weng, T. S. Huang, and N. Ahuja, "Motion and structure from line correspondences: Closed form solution and uniqueness results," *IEEE Trans. Pattern Analysis and Machine Intell.*, vol. 14, no. 3, pp. 318–336, 1992.

[14] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization method," *Int. J. of Computer Vision*, vol. 9, no. 2, pp. 137–154, 1992.

[15] C. J. Poelman and T. Kanade, "A paraperspective factorization method for shape and motion recovery," in *Proc. 3rd European Conference on Computer Vision*, (Stockholm, Sweden), pp. 97–108, Springer-Verlag, May 1994.

[16] L. Quan and T. Kanade, "Affine structure from line correspondences with uncalibrated affine cameras," *IEEE Trans. Pattern Analysis and Machine Intell.*, vol. 19, no. 8, pp. 834–845, 1997.

[17] P. Sturm and B. Triggs, "A factorization based algorithm for multi-image projective structure and motion," in *Proc. 4th European Conference on Computer Vision*, vol. 1064, (Cambridge, UK), pp. 709–720, Springer Verlag, Berlin, Apr. 1996.

[18] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, pp. 133–135, 1981.

[19] O. Faugeras, "What can be seen in three dimensions with a stereo rig?," in *Proc. 2nd European Conf. on Computer Vision* (G. Sandini, ed.), vol. 588 of *Lecture Notes in Computer Science*, (Santa Margherita Ligure, Italy), pp. 563–578, Springer-Verlag, May. 1992.

[20] G. Xu and Z. Zhang, eds., *Epipolar Geometry in Stereo, Motion and Object Recognition: A Unified Approach.* Series in Mathematical Imaging and Vision, Dordrecht, Netherlands: Kluwer Academic Publishers, 1997.

[21] A. Shashua, “Algebraic functions for recognition,” *IEEE Trans. Pattern Analysis and Machine Intell.*, vol. 17, no. 8, pp. 779–789, 1995.

[22] R. Hartley, “A linear method for reconstruction from points and lines,” in *Proc. 5th International Conference on Computer Vision*, (Cambridge, MA), pp. 882–887, June 1995.

[23] A. Heyden, “Reconstruction from image sequences by means of relative depth,” in *Proc. 5th International Conference on Computer Vision*, (Cambridge, MA), pp. 57–66, June 1995.

[24] P. Beardsley, P. Torr, and A. Zisserman, “3D model acquistions from extended image sequences,” in *Proc. 4th European Conference on Computer Vision*, (Cambridge, UK), pp. 683–695, Springer Verlag, Berlin, Apr. 1996.

[25] O. Faugeras and B. Mourrain, “On the geometry and algebra of the point and line correspondences between N images,” in *Proc. 5th International Conference on Computer Vision*, (Cambridge, MA), pp. 951–956, June 1995.

[26] A. Heyden, G. Sparr, and K. Åström, “Perception and action using multilinear forms,” in *Proc. AFPAC’97: Algebraic Frames for the Perception-Action Cycle* (G. Sommer and J. J. Koenderink, eds.), (Kiel, Germany), pp. 54–65, Springer Verlag, Berlin, Sept. 1997.

[27] T. Lindeberg and L. Bretzner, “Förfarande och anordning för överföring av information genom rörelsedetektering, samt användning av anordningen.” Patent pending 9800884-0., 1998.

[28] R. Cipolla, Y. Okamoto, and Y. Kuno, “Robust structure from motion using motion parallax,” in *Proc. 4th Int. Conf. on Computer Vision*, (Berlin, Germany), pp. 374–382, May. 1993.

[29] W. T. Freeman and C. D. Weissman, “Television control by hand gestures,” in *Proc. Int. Conf. on Face and Gesture Recognition*, (Zurich, Switzerland), 1995.

[30] C. Maggioni and B. Kämmerer, “Gesturecomputer-history, design and applications,” in *Computer vision for human-computer interaction* (R. Cipolla and A. Pentland, eds.), (Cambridge, U.K.), pp. 23–52, Cambridge University Press, 1998.

[31] J. J. Kuch and T. S. Huang, “Vision based hand modelling and tracking for virtual teleconferencing and telecollaboration,” in *Proc. 5th International Conference on Computer Vision*, (Cambridge, MA), pp. 666–671, June 1995.

[32] J. Lee and T. L. Kunii, “Model-based analysis of hand posture,” *Computer Graphics and Applications*, pp. 77–86, Jun. 1995.

[33] T. Heap and D. Hogg, “Towards 3D hand tracking using a deformable model,” in *Int. Conf. on Automatic Face and Gesture Recognition*, (Killington, Vermont), pp. 140–145, Oct. 1996.

[34] Y. Yasumuro, Q. Chen, and K. Chihara, “Three-dimensional modelling of the human hand with motion constraints,” *Image and Vision Computing*, vol. 17, pp. 149–156, Mar. 1999.

[35] L. Bretzner and T. Lindeberg, “Feature tracking with automatic selection of spatial scales,” *Computer Vision and Image Understanding*, vol. 71, pp. 385–392, Sep. 1998.

[36] L. Bretzner and T. Lindeberg, “On the handling of spatial and temporal scales in feature tracking,” in *Scale-Space Theory in Computer Vision: Proc. First Int. Conf. Scale-Space’97* (B. M. ter Haar Romeny, L. M. J. Florack, J. J. Koenderink, and M. A. Viergever, eds.), vol. 1252 of *Lecture Notes in Computer Science*, (Utrecht, The Netherlands), pp. 128–139, Springer Verlag, New York, July 1997.

[37] J. J. Koenderink, “The structure of images,” *Biological Cybernetics*, vol. 50, pp. 363–370, 1984.

[38] T. Lindeberg, *Scale-Space Theory in Computer Vision.* The Kluwer International Series in Engineering and Computer Science, Dordrecht, Netherlands: Kluwer Academic Publishers, 1994.

[39] T. Lindeberg, “Feature detection with automatic scale selection,” *Int. J. of Computer Vision*, vol. 30, no. 2, pp. 77–116, 1998.

[40] T. Lindeberg, “Edge detection and ridge detection with automatic scale selection,” *Int. J. of Computer Vision*, vol. 30, no. 2, pp. 117–154, 1998.

[41] L. Bretzner and T.Lindeberg, “Qualitative multi-scale feature hierarchies for object tracking,” in *Proc. 2nd International Conference on Scale-Space Theories in Computer Vision* (O. F. O. M. Nielsen, P. Johansen and J. Weickert, eds.), vol. 1682, (Corfu, Greece), pp. 117–128, Springer Verlag, Sep. 1999. Lecture Notes in Computer Science, (Extended version available as Tech. Rep. ISRN KTH/NA/P–99/09–SE).

[42] A. Shashua, "Trilinear tensor: The fundamental construct of multiple-view geometry and its applications," in *Proc. AFPAC'97: Algebraic Frames for the Perception-Action Cycle* (G. Sommer and J. J. Koenderink, eds.), (Kiel, Germany), pp. 190–206, Springer Verlag, Berlin, Sept. 1997.

[43] S. Ullman and R. Basri, "Recognition by linear combinations of models," *IEEE Trans. Pattern Analysis and Machine Intell.*, vol. 13, no. 10, pp. 992–1006, 1991.

[44] L. S. Shapiro, *Affine analysis of image sequences.* Cambridge, England: Cambridge University Press, 1995.

[45] K. Kanatani, "Factorization without factorization: Statistical analysis," tech. rep., Dept. of Computer Science, Gunma University, Japan, 1998.

[46] M. Mühlich and R. Mester, "The role of total least squares in motion analysis," in *Proc. 5th European Conference on Computer Vision* (H. Burkhardt and B. Neumann, eds.), (Freiburg, Germany), Springer Verlag, Berlin, 1998.

[47] L. Bretzner and T. Lindeberg, "Use your hand as a 3-D mouse or relative orientation from extended sequences of sparse point and line correspondances using the affine trifocal tensor," in *Proc. 5th European Conference on Computer Vision* (H. Burkhardt and B. Neumann, eds.), vol. 1406 of *Lecture Notes in Computer Science*, (Freiburg, Germany), pp. 141–157, Springer Verlag, Berlin, June 1998.

[48] F. Kahl and A. Heyden, "Structure and motion from points, lines and conics with affine cameras," in *Proc. 5th European Conference on Computer Vision* (H. Burkhardt and B. Neumann, eds.), vol. 1406 of *Lecture Notes in Computer Science*, (Freiburg, Germany), pp. 327–341, Springer Verlag, Berlin, June 1998.

[49] L. Quan, Y. Ohta, and R. Mohr, "Geometry of multiple affine views," in *3D Structure from Multiple Images of Large-Scale Environments* (R. Koch and L. van Gool, eds.), vol. 1506 of *Lecture Notes in Computer Science*, (Freiburg, Germany), pp. 32–46, Springer Verlag, Berlin, June 1998.

[50] P. Mendonca and R. Cipolla, "Analysis and computation of an affine trifocal tensor," in *Proc. British Machine Vision Conference*, (Southampton), pp. 141–157, 1998.

[51] T. Thórhallsson and D. W. Murray, "The tensors of three affine views," in *Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition*, (Fort Collins, Colorado), pp. 450–456, Jun. 1999.

[52] E. Hayman, T. Thorhallsson, and D. W. Murray, "Zoom-invariant tracking using point and lines in affine views, an application of the affine multifocal tensors," in *Proc. 7th Int. Conf. on Computer Vision*, (Corfu, Greece), pp. 269–276, IEEE Computer Society Press, 1999.

[53] F. Kahl and A. Heyden, "Robust self-calibration and euclidean reconstruction via affine approximation," in *International Conference on Pattern Recognition*, (Brisbane, Australia), pp. 47–55, 1998.