Use Your Hand as a 3-D Mouse, or, Relative Orientation from Extended Sequences of Sparse Point and Line Correspondences Using the Affine Trifocal Tensor*

Lars Bretzner and Tony Lindeberg

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

Abstract. This paper addresses the problem of computing three-dimensional structure and motion from an unknown rigid configuration of point 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 also developed to handle point features and line features simultaneously in image sequences. This theory is applied to a specific problem in human-computer interaction of capturing three-dimensional rotations from gestures of a human hand. Besides the obvious application, this test problem 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.

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

* The support from the Swedish Research Council for Engineering Sciences, TFR, is gratefully acknowledged. Email: bretzner@nada.kth.se, tony@nada.kth.se

1979, Maybank 1992, Huang & Lee 1989, Huang & Netravali 1994) and others. With the introduction of the affine camera model (Koenderink & van Doorn 1991, Mundy & Zisserman 1992) a large number of approaches have been developed, including (Shapiro 1995, Beardsley et al. 1994, McLauchlan et al. 1994, Torr 1995) to mention just a few. Line correspondences have been studied by (Spetsakis & Aloimonos 1990, Weng et al. 1992), and factorization methods for points and lines constitute a particularly interesting development (Tomasi & Kanade 1992, Morita & Kanade 1997, Quan & Kanade 1997, Sturm & Triggs 1996). These directions of research have recently been combined with the ideas behind the fundamental matrix (Longuet-Higgins 1981, Faugeras 1992, Xu & Zhang 1997) and have lead to the trilinear tensor (Shashua 1995, Hartley 1995, Heyden 1995) as a unified model for point and line correspondences for three cameras, with interesting applications (Beardsley et al. 1996) as well as a deeper understanding of the relations between point features and line features over multiple views (Faugeras & Mourrain 1995, Heyden et al. 1997).

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. It will also be shown that motion estimation from this tensor is more straightforward.

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 straightforward man-machine interface for 3-D orientation interaction (Lindeberg & Bretzner 1998) can be designed based on the theory presented and using no other user equipment than the operator's own hand. For more details, see (Bretzner & Lindeberg 1998).

2 Geometric problem and extraction of image features

A specific application we are interested in is to measure changes in the orientation of a human hand, as a straightforward interface to transfer 3-D rotational information to a computer using no other user equipment than the operator's own hand. In contrast to previous approaches for human–computer interaction that are based on detailed geometric hand models (such as (Lee & Kunii 1995, Heap & Hogg 1996)) we shall here explore a model based on qualitative features only. This model involves the thumb, the index finger and the middle finger, 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 three point correspondences and three line correspondences, and the task is to compute changes in the 3-D orientation of such a configuration, which is assumed to be rigid. It is worth noting that neither the trajectories of

point features or line features *per se* are sufficient to compute the motion information we are interested in. The problem requires the combination of point and line features. 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 & Lindeberg 1996, Bretzner & Lindeberg 1997). Blob features corresponding to the finger tips are computed from points $(x, y;\ t)$ in scale-space (Koenderink 1984, Lindeberg 1994) 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 1994). 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 defined by (Lindeberg 1996)

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

where L_{pp} and L_{qq} are the eigenvalues of the Hessian matrix and the normalization parameter $\gamma = 0.875$. At each ridge feature, a windowed second moment matrix (Förstner & Gülch 1987, Bigün et al. 1991, Lindeberg 1994)

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

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.

The left column in figure 3 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 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 1995, Shashua 1997, Hartley 1995, Heyden et al. 1997).

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 & Kanade 1992, Ullman & Basri 1991). Frameworks for capturing line correspondences over multiple affine views have been presented by (Quan & Kanade 1997) and for point features under perspective projection by (Sturm & Triggs 1996).

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'':

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

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

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

Following (Faugeras & Mourrain 1995) and (Shashua 1997), let us 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 throughout 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 1995, Hartley 1995)

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

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 relations between the projections p, p' and p'' of P (Shashua 1997):

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

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

$$\begin{aligned}
L^T &= l^T M = (l_1,\, l_2,\, l_3\, 0),\\
L'^T &= l'^T M' = (l'_j\, a_1^j,\, l'_j\, a_2^j,\, l'_j\, a_3^j,\, l'_j\, u'^j),\\
L''^T &= l''^T M'' = (l''_k\, b_1^k,\, l''_k\, b_2^k,\, l''_k\, b_3^k,\, l''_k\, u''^k).
\end{aligned} \tag{12}$$

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

$$\operatorname{rank} \begin{pmatrix} l_1 & l'_j\, a_1^j & l''_k\, b_1^k \\ l_2 & l'_j\, a_2^j & l''_k\, b_2^k \\ l_3 & l'_j\, a_3^j & l''_k\, b_3^k \\ 0 & l'_j\, u'^j & l''_k\, u''^k \end{pmatrix} = 2. \tag{13}$$

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:

$$\begin{aligned}
(l_2 T_3^{jk} - l_3 T_2^{jk})\, l'_j\, l''_k &= 0,\\
(l_1 T_3^{jk} - l_3 T_1^{jk})\, l'_j\, l''_k &= 0,\\
(l_1 T_2^{jk} - l_2 T_1^{jk})\, l'_j\, l''_k &= 0.
\end{aligned} \tag{14}$$

These expressions provide a compact characterization of the trilinear line relations first introduced by (Spetsakis & Aloimonos 1990).

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 1995, Hartley 1995).

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:

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

$$q' = \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = M'Q = \begin{pmatrix} c_1^1 & c_2^1 & c_3^1 & v'^1 \\ c_1^2 & c_2^2 & c_3^2 & v'^2 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ \lambda \\ 1 \end{pmatrix}, \tag{16}$$

$$q'' = \begin{pmatrix} x'' \\ y'' \\ 1 \end{pmatrix} = M''Q = \begin{pmatrix} d_1^1 & d_2^1 & d_3^1 & v''^1 \\ d_1^2 & d_2^2 & d_3^2 & v''^2 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ \lambda \\ 1 \end{pmatrix}. \tag{17}$$

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 between the image coordinates of q, q' and q'':

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

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

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

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 1995) 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 $4K + 2L \geq 15$.

3.3 The centered affine camera and its relations to perspective

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. Then, the expressions (18) and (20) for the point and line correspondences reduce to:

$$\begin{aligned} (c_3^1 d_1^1 - c_1^1 d_3^1)x + (c_3^1 d_2^1 - c_2^1 d_3^1)y + d_3^1 x' - c_3^1 x'' &= 0, \\ (c_3^2 d_1^1 - c_1^2 d_3^1)x + (c_3^2 d_2^1 - c_2^2 d_3^1)y + d_3^1 y' - c_3^2 x'' &= 0, \\ (c_3^1 d_1^2 - c_1^1 d_3^2)x + (c_3^1 d_2^2 - c_2^1 d_3^2)y + d_3^2 x' - c_3^1 y'' &= 0, \\ (c_3^2 d_1^2 - c_1^2 d_3^2)x + (c_3^2 d_2^2 - c_2^2 d_3^2)y + d_3^2 y' - c_3^2 y'' &= 0, \\ l_1(l'_j c_3^j)l''_3 - l_1 l'_3(l''_k d_3^k) + l_3(l'_j c_1^j)(l''_k d_3^k) - l_3(l'_j c_3^j)(l''_k d_1^k) &= 0, \\ l_2(l'_j c_3^j)l''_3 - l_2 l'_3(l''_k d_3^k) + l_3(l'_j c_2^j)(l''_k d_3^k) - l_3(l'_j c_3^j)(l''_k d_2^k) &= 0, \\ l_1(l'_j c_2^j)(l''_k d_3^k) - l_1(l'_j c_3^j)(l''_k d_2^k) + l_2(l'_j c_3^j)(l''_k d_1^k) - (l'_j c_1^j)(l''_k d_3^k) &= 0. \end{aligned} \tag{21}$$

Structurally, there is a strong similarity between these relationships for the centered 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:

$$Q = (x, y, \lambda, 1)^T \quad \Rightarrow \quad P = (x, y, 1, \lambda)^T, \tag{22}$$

- Interchange columns 3 and 4 in the camera matrices:

$$\begin{pmatrix} a_1^1 & a_2^1 & a_3^1 & u'^1 \\ a_1^2 & a_2^2 & a_3^2 & u'^2 \\ a_1^3 & a_2^3 & a_3^3 & u'^3 \end{pmatrix} = \begin{pmatrix} c_1^1 & c_2^1 & 0 & c_3^1 \\ c_1^2 & c_2^2 & 0 & c_3^2 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} b_1^1 & b_2^1 & b_3^1 & u''^1 \\ b_1^2 & b_2^2 & b_3^2 & u''^2 \\ b_1^3 & b_2^3 & b_3^3 & u''^3 \end{pmatrix} = \begin{pmatrix} d_1^1 & d_2^1 & 0 & d_3^1 \\ d_1^2 & d_2^2 & 0 & d_3^2 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{23}$$

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 in (21) for three centered cameras can be expressed on the form (11) and (14) with the *centered affine trifocal tensor* defined by

$$\mathcal{T}_i^{jk} = a_i^j u''^k - b_i^k u'^j = \{(23)\} = c_i^j d_3^k - d_i^k c_3^j. \tag{24}$$

Written out explicitly, the components of $\mathcal{T}_i^{jk}$ are given by

$$\begin{aligned}
\mathcal{T}_1^{11} &= c_1^1 d_3^1 - d_1^1 c_3^1, & \mathcal{T}_1^{12} &= c_1^1 d_3^2 - d_1^2 c_3^1, & \mathcal{T}_1^{13} &= c_1^1 d_3^3 - d_1^3 c_3^1 &&= 0,\\
\mathcal{T}_1^{21} &= c_1^2 d_3^1 - d_1^1 c_3^2, & \mathcal{T}_1^{22} &= c_1^2 d_3^2 - d_1^2 c_3^2, & \mathcal{T}_1^{23} &= c_1^2 d_3^3 - d_1^3 c_3^2 &&= 0,\\
\mathcal{T}_1^{31} &= c_1^3 d_3^1 - d_1^1 c_3^3 &&= 0, & \mathcal{T}_1^{32} &= c_1^3 d_3^2 - d_1^2 c_3^3 &&= 0, & \mathcal{T}_1^{33} &= c_1^3 d_3^3 - d_1^3 c_3^3 &&= 0,\\
\mathcal{T}_2^{11} &= c_1^1 d_3^1 - d_1^1 c_3^1, & \mathcal{T}_2^{12} &= c_1^1 d_3^2 - d_1^2 c_3^1, & \mathcal{T}_2^{13} &= c_1^1 d_3^3 - d_1^3 c_3^1 &&= 0,\\
\mathcal{T}_2^{21} &= c_1^2 d_3^1 - d_1^1 c_3^2, & \mathcal{T}_2^{22} &= c_1^2 d_3^2 - d_1^2 c_3^2, & \mathcal{T}_2^{23} &= c_1^2 d_3^3 - d_1^3 c_3^2 &&= 0,\\
\mathcal{T}_2^{31} &= c_1^3 d_3^1 - d_1^1 c_3^3 &&= 0, & \mathcal{T}_2^{32} &= c_1^3 d_3^2 - d_1^2 c_3^3 &&= 0, & \mathcal{T}_2^{33} &= c_1^3 d_3^3 - d_1^3 c_3^3 &&= 0,\\
\mathcal{T}_3^{11} &= c_1^1 d_3^1 - d_1^1 c_3^1 &&= 0, & \mathcal{T}_3^{12} &= c_1^1 d_3^2 - d_1^2 c_3^1 &&= 0, & \mathcal{T}_3^{13} &= c_1^1 d_3^3 - d_1^3 c_3^1 &&= -c_3^1,\\
\mathcal{T}_3^{21} &= c_1^2 d_3^1 - d_1^1 c_3^2 &&= 0, & \mathcal{T}_3^{22} &= c_1^2 d_3^2 - d_1^2 c_3^2 &&= 0, & \mathcal{T}_3^{23} &= c_1^2 d_3^3 - d_1^3 c_3^2 &&= -c_3^2,\\
\mathcal{T}_3^{31} &= c_1^3 d_3^1 - d_1^1 c_3^3 &&= d_3^1, & \mathcal{T}_3^{32} &= c_1^3 d_3^2 - d_1^2 c_3^3 &&= d_3^2, & \mathcal{T}_3^{33} &= c_1^3 d_3^3 - d_1^3 c_3^3 &&= 0,
\end{aligned} \tag{25}$$

and the relations between point and line correspondences in (21) can be written

$$\begin{aligned}
\mathcal{T}_3^{13} x'' + \mathcal{T}_3^{31} x' - \mathcal{T}_1^{11} x - \mathcal{T}_2^{11} y &= 0,\\
\mathcal{T}_3^{13} y'' + \mathcal{T}_3^{32} x' - \mathcal{T}_1^{12} x - \mathcal{T}_2^{12} y &= 0,\\
\mathcal{T}_3^{23} x'' + \mathcal{T}_3^{31} y' - \mathcal{T}_1^{21} x - \mathcal{T}_2^{21} y &= 0,\\
\mathcal{T}_3^{23} y'' + \mathcal{T}_3^{32} y' - \mathcal{T}_1^{22} x - \mathcal{T}_2^{22} y &= 0,
\end{aligned} \tag{26}$$

$$\begin{aligned}
&l_3(l_1' l_1'' \mathcal{T}_1^{11} + l_1' l_2'' \mathcal{T}_1^{12} + l_2' l_1'' \mathcal{T}_1^{21} + l_2' l_2'' \mathcal{T}_1^{22})\\
&- l_1(l_1' l_3'' \mathcal{T}_3^{13} + l_2' l_3'' \mathcal{T}_3^{23} + l_3' l_1'' \mathcal{T}_3^{31} + l_3' l_2'' \mathcal{T}_3^{32}) = 0,\\
&l_3(l_1' l_1'' \mathcal{T}_2^{11} + l_1' l_2'' \mathcal{T}_2^{12} + l_2' l_1'' \mathcal{T}_2^{21} + l_2' l_2'' \mathcal{T}_2^{22})\\
&- l_2(l_1' l_3'' \mathcal{T}_3^{13} + l_2' l_3'' \mathcal{T}_3^{23} + l_3' l_1'' \mathcal{T}_3^{31} + l_3' l_2'' \mathcal{T}_3^{32}) = 0.
\end{aligned} \tag{27}$$

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 $4(K-1) + 2L \geq 11$.

4 Orientation from the centered affine trifocal tensor

To compute the camera parameters from the affine trifocal tensor, we largely follow the approach that (Hartley 1995) uses for three perspective cameras. The

calculations can, however, be simplified with affine cameras. From (25) we directly get

$$c_3^1 = -\mathcal{T}_3^{13}, \qquad d_3^1 = \mathcal{T}_3^{31}, \qquad c_3^2 = -\mathcal{T}_3^{23}, \qquad d_3^2 = \mathcal{T}_3^{32}. \tag{28}$$

Given these c_3^j and d_3^k, the remaining c_i^j and d_i^k can be computed from (25) using

$$\begin{pmatrix} d_3^1 & & -c_3^1 & & & \\ d_3^2 & & & -c_3^1 & & \\ & d_3^1 & -c_3^2 & & & \\ & d_3^2 & & -c_3^2 & & \\ & & d_3^1 & & -c_3^1 & \\ & & d_3^2 & & & -c_3^1 \\ & & & d_3^1 & -c_3^2 & \\ & & & d_3^2 & & -c_3^2 \end{pmatrix} \begin{pmatrix} c_1^1 \\ c_1^2 \\ c_2^1 \\ c_2^2 \\ d_1^1 \\ d_1^2 \\ d_2^1 \\ d_2^2 \end{pmatrix} = \begin{pmatrix} \mathcal{T}_1^{11} \\ \mathcal{T}_1^{12} \\ \mathcal{T}_1^{21} \\ \mathcal{T}_1^{22} \\ \mathcal{T}_2^{11} \\ \mathcal{T}_2^{12} \\ \mathcal{T}_2^{21} \\ \mathcal{T}_2^{22} \end{pmatrix}. \tag{29}$$

The camera matrices are, however, not uniquely determined. The centered affine trifocal tensor $\mathcal{T}_i^{jk}$ in (24) is invariant under transformations of the type

$$\tilde{c}_i^j = c_i^j + \gamma_i\, c_3^j, \qquad \tilde{d}_i^k = d_i^k + \gamma_i\, d_3^k. \tag{30}$$

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)

$$\tilde{N}' = \begin{pmatrix} \tilde{c}_1^1 & \tilde{c}_2^1 & \tilde{c}_3^1 \\ \tilde{c}_1^2 & \tilde{c}_2^2 & \tilde{c}_3^2 \end{pmatrix} = \begin{pmatrix} c_1^1 & c_2^1 & c_3^1 \\ c_1^2 & c_2^2 & c_3^2 \end{pmatrix} \begin{pmatrix} 1 & & \\ & 1 & \\ \gamma_1 & \gamma_2 & \gamma_3 \end{pmatrix} = N'\,\Gamma. \tag{31}$$

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)

$$\tilde{N}' = \sigma' \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} R' = \sigma' \begin{pmatrix} \rho'^1_1 & \rho'^1_2 & \rho'^1_3 \\ \rho'^2_1 & \rho'^2_2 & \rho'^2_3 \end{pmatrix}, \tag{32}$$

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

$$\tilde{N}'{\tilde{N}'}^T = N'\Gamma\Gamma^T{N'}^T = (\sigma')^2 I_{2\times 2}, \tag{33}$$

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

$$\Gamma\Gamma^T = \begin{pmatrix} 1 & 0 & \gamma_1 \\ 1 & 0 & \gamma_2 \\ \gamma_1 & \gamma_2 & \gamma_1^2 + \gamma_2^2 + \gamma_3^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \xi \\ 0 & 1 & \eta \\ \xi & \eta & \zeta \end{pmatrix} \tag{34}$$

we rewrite (33) as

$$\begin{pmatrix} 2c_1^1c_3^1 & 2c_2^1c_3^1 & (c_3^1)^2 & -1 & 0 \\ c_1^1c_3^2+c_3^1c_1^2 & c_2^1c_3^2+c_3^1c_2^2 & c_3^1c_3^2 & 0 & 0 \\ 2c_1^2c_3^2 & 2c_2^2c_3^2 & (c_3^2)^2 & -1 & 0 \\ 2d_1^1d_3^1 & 2d_2^1d_3^1 & (d_3^1)^2 & 0 & -1 \\ d_1^1d_3^2+d_3^1d_1^2 & d_2^1d_3^2+d_3^1d_2^2 & d_3^1d_3^2 & 0 & 0 \\ 2d_1^2d_3^2 & 2d_2^2d_3^2 & (d_3^2)^2 & 0 & -1 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \\ \zeta \\ (\sigma')^2 \\ (\sigma'')^2 \end{pmatrix} = - \begin{pmatrix} (c_1^1)^2+(c_2^1)^2 \\ c_1^1c_1^2+c_2^1c_2^2 \\ (c_1^2)^2+(c_2^2)^2 \\ (d_1^1)^2+(d_2^1)^2 \\ d_1^1d_1^2+d_2^1d_2^2 \\ (d_1^2)^2+(d_2^2)^2 \end{pmatrix}. \tag{35}$$

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 (28) and (29). Then, Γ is given by

$$\Gamma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \xi & \eta & \pm\sqrt{\zeta - \xi^2 - \eta^2} \end{pmatrix}, \tag{36}$$

and we estimate the first two rows of R' in (32) 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:

$$\rho = \begin{pmatrix} \rho_1^1 & \rho_2^1 & \rho_3^1 \\ \rho_1^2 & \rho_2^2 & \rho_3^2 \\ \rho_1^3 & \rho_2^3 & \rho_3^3 \end{pmatrix}, \qquad \bar{\rho} = \begin{pmatrix} \rho_1^1 & \rho_2^1 & -\rho_3^1 \\ \rho_1^2 & \rho_2^2 & -\rho_3^2 \\ -\rho_1^3 & -\rho_2^3 & \rho_3^3 \end{pmatrix}. \tag{37}$$

This ambiguity reflects the fact that for scaled orthographic projection we cannot distinguish between positive rotation of a point in front of the center of rotation and 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.

The matrices obtained from (37) 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.

5 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 & Kanade 1992, Quan & Kanade 1997, Sturm & Triggs 1996). 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 & Kanade 1992). 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 notion (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

$$\begin{pmatrix} x_k^n \\ y_k^n \end{pmatrix} = M^n P_k = \begin{pmatrix} - \alpha^{nT} - \\ - \beta^{nT} - \end{pmatrix} \begin{pmatrix} X_k \\ Y_k \\ Z_k \end{pmatrix}, \tag{38}$$

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

$$\lambda_l^n \begin{pmatrix} u_l^n \\ v_k^n \end{pmatrix} = M^n Q_l = \begin{pmatrix} - \alpha^{nT} - \\ - \beta^{nT} - \end{pmatrix} \begin{pmatrix} U_l \\ V_l \\ W_l \end{pmatrix}, \tag{39}$$

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 correspondences and L line correspondences over N image frames, we model these measurements together by a matrix G

$$\begin{pmatrix} x_1^1 & \dots & x_K^1 & \lambda_1^1 u_1^1 & \dots & \lambda_L^1 u_L^1 \\ y_1^1 & \dots & y_K^1 & \lambda_1^1 v_1^1 & \dots & \lambda_L^1 v_L^1 \\ \vdots & & \vdots & \vdots & & \vdots \\ x_1^N & \dots & x_K^N & \lambda_1^N u_1^N & \dots & \lambda_L^N u_L^N \\ y_1^N & \dots & y_K^N & \lambda_1^N v_1^N & \dots & \lambda_L^N v_L^N \end{pmatrix} = \begin{pmatrix} - \alpha^{1T} - \\ - \beta^{1T} - \\ \vdots \\ - \alpha^{NT} - \\ - \beta^{NT} - \end{pmatrix} \begin{pmatrix} X_1 & \dots & X_K & U_1 & \dots & U_L \\ Y_1 & \dots & Y_K & V_1 & \dots & V_L \\ Z_1 & \dots & Z_K & W_1 & \dots & W_L \end{pmatrix} \tag{40}$$

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

$$\begin{vmatrix} x_k^n & x_{k'}^n & x_{k''}^n & \lambda_l^n u_l^n \\ y_k^n & y_{k'}^n & y_{k''}^n & \lambda_l^n v_l^n \\ x_k^{n'} & x_{k'}^{n'} & x_{k''}^{n'} & \lambda_l^{n'} u_l^{n'} \\ y_k^{n'} & y_{k'}^{n'} & x_{k''}^{n'} & \lambda_l^{n'} v_l^{n'} \end{vmatrix} = 0. \tag{41}$$

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 & Kanade 1997) and to point features by (Sturm & Triggs 1996).

Given only three points, however, as in our test problem, 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).

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 & Kanade 1997) let us for each such triplet $n, n', n'' \in [1..N]$, insert the following shape matrix

$$\begin{pmatrix} X_1 & X_2 & X_3 \\ Y_1 & Y_2 & Y_3 \\ Z_1 & Z_2 & Z_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{42}$$

into the projection equation (40) for $K = 3$ point features:

$$H^{n,n',n''} = \begin{pmatrix} -\alpha^{nT} & - \lambda_1^n u_1^n & \dots & \lambda_L^n u_L^n \\ -\beta^{nT} & - \lambda_1^n v_1^n & \dots & \lambda_L^n v_L^n \\ -\alpha^{n'T} & - \lambda_1^{n'} u_1^{n'} & \dots & \lambda_L^{n'} u_L^{n'} \\ -\beta^{n'T} & - \lambda_1^{n'} v_1^{n'} & \dots & \lambda_L^{n} v_L^{n'} \\ -\alpha^{n''T} & - \lambda_1^{n''} u_1^{n''} & \dots & \lambda_L^{n''} u_L^{n''} \\ -\beta^{n''T} & - \lambda_1^{n''} v_1^{n''} & \dots & \lambda_L^{n} v_L^{n''} \end{pmatrix}. \tag{43}$$

Then, since the rank of the right hand side in (40) 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 4.

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:

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

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 from 2 to 4). Moreover, a ranking of the image triplets is carried out based on 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:

$$D_l = U_l \Sigma_l V_l^T \quad \Rightarrow \quad \Lambda_l = \text{the last row of } V_l. \tag{44}$$

The λ_l^n values are inserted into G in (40) and a singular value decomposition is computed $G = U_G \Sigma_G V_G^T$. All elements except the three first ones in Σ_G are set to zero to reduce the rank to three, and 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 & Kanade 1992, Quan & Kanade 1997).

6 Experiments

To investigate the properties of this framework for computing relative orientation, let us first apply it to synthetic data, which will be generated by the following procedure: A three-dimensional point and line model is generated from three lines and three points with the approximate shape of the thumb, the index finger and the middle finger of a hand. This configuration is subjected to a smooth rotation, where the mean of the three directional vectors rotates by $\Delta\phi \approx 2$ degrees between each frame, while the configuration also rotates by $\Delta\psi \approx 2$ degrees per frame around this moving axis. For each frame, an affine projection is computed involving variations in scale and translation. White Gaussian noise with zero mean and standard deviation Σ is added to the image projections, where s is determined from a noise level ν according to $s = \nu D/2$ and D is the diameter of the circle that interpolates the three (undistorted) image points.

The difference between the orientation estimates and the true orientation is measured in the following ways: (i) the Frobenius norm of the difference between the true and the estimated rotation matrix, (ii) two geometric angles defined as follows: θ is the angle between the true and the estimate rotation axis (this rotation axis is the real eigenvector of the rotation matrix), and ϕ is the difference between the estimated and the real rotation around this rotation axis.

Figure 1 shows the result of estimating rotations from these point and line features using the affine trifocal tensor applied to triplets of frames only. The distance between adjacent frames varies from $\Delta n = 2$ to 20 frames, and all results are average values over 10 experiments.

Figure 2 shows a corresponding evaluation of the joint factorization approach. Here, the error is shown as function of the number of frames for computing the

Noise level	Error measure in 3-D		
$\nu = 0.002$	Frobenius	θ	ϕ
$\Delta n = 2$	0.22	31.00	5.92
$\Delta n = 5$	0.39	20.35	13.15
$\Delta n = 10$	0.11	2.69	3.35
$\Delta n = 20$	0.02	0.46	0.30

Noise level	Error measure in 3-D		
$\nu = 0.005$	Frobenius	θ	ϕ
$\Delta n = 2$	0.27	48.74	6.18
$\Delta n = 5$	0.50	35.33	17.18
$\Delta n = 10$	0.29	11.58	8.85
$\Delta n = 20$	0.08	1.58	1.30

Fig. 1: Experimental evaluation of the noise sensitivity when computing 3-D orientation estimates by determining the affine trifocal tensor from three point correspondences and three line correspondences over three affine views. From left to right, the tables show the following error measures: (left) the Frobenius norm, (middle) the rotation axis θ, and (right) the rotation angle ϕ. The results are shown for two noise levels.

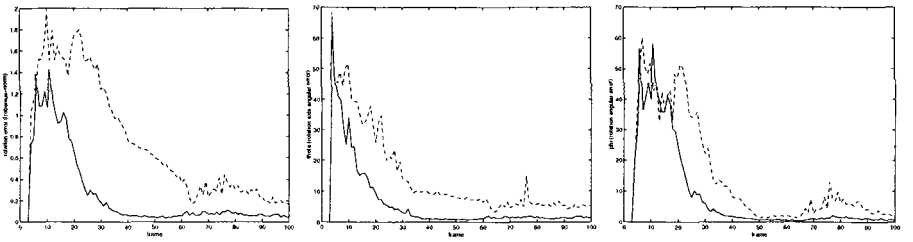

Fig. 2: Experimental evaluation of the noise sensitivity when computing 3-D orientation using the joint factorization of point features and line features. From left to right, the graphs how the following error measures depend on the number of frames: (left) the Frobenius norm, (middle) the rotation axis θ, and (right) the rotation angle ϕ. Three curves are shown in each graph, for noise levels $\nu = 0.002$ and 0.005, respectively.

factorization. (The results are averages over 8 experiments with incremental computations.) As expected, the error decreases with the number of image frames.

Finally, figure 3 shows the result of computing corresponding estimates by applying the joint factorization approach to the feature trajectories obtained by tracking blob and ridge features of a human hand as described in section 2. Here, since no ground truth is available, the result is illustrated by subjecting a synthetic cube to the rotation estimates computed from the feature trajectories.

7 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 (Koenderink & van Doorn 1991, Mundy & Zisserman 1992, Shapiro 1995, Faugeras 1995), perspective point correspondences (Ullman 1979, Huang & Netravali 1994) and line correspondences (Spetsakis & Aloimonos 1990, Weng et al. 1992) as can be modelled by the trilinear tensor (Shashua 1995, Hartley 1995, Faugeras & Mourrain 1995, Shashua 1997). It also builds upon factorization approaches for affine (Tomasi & Kanade 1992, Quan & Kanade 1997, Morita & Kanade 1997) and perspective (Sturm & Triggs 1996) projection.

We propose that the (centered) affine trifocal tensor constitutes a canonical tool to model point and line correspondences in triplets of affine views (section 3). This extends the advances by (Torr 1995) as well as the abovementioned 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 4) is also more straightforward.

To capture point and line correspondences in dense time sequences, we have also applied a factorization approach (section 5), to which the affine trifocal

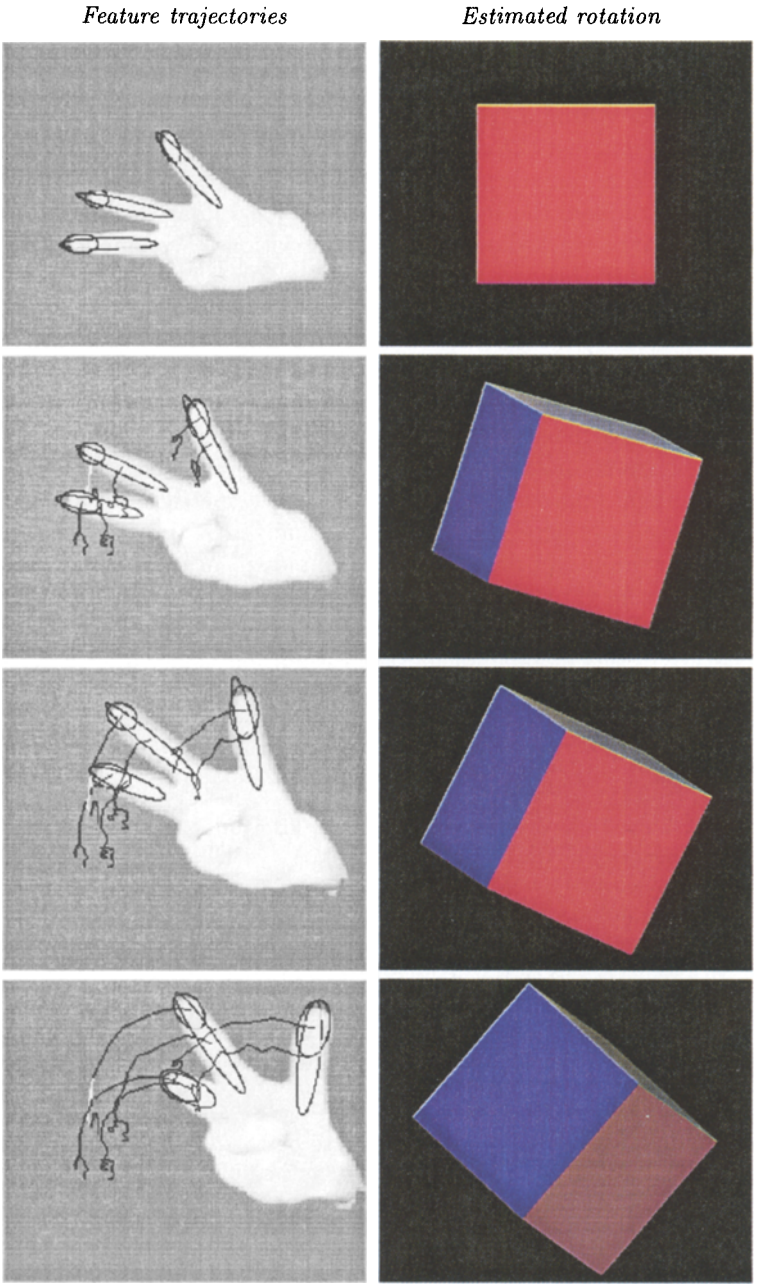

Fig. 3: Estimates of relative orientation from hand gestures. The left column shows point features and line features obtained from a feature tracker with automatic scale selection (Bretzner & Lindeberg 1997). The right column shows the result of computing changes in 3-D orientation using the joint factorization of point and line features in section 5. The results are illustrated by subjecting a three-dimensional cube to the estimated rotations.

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 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 (Lindeberg & Bretzner 1998). Contrary to the more common approach of using detailed geometric hand models (Lee & Kunii 1995), 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.

References

Beardsley, P., Torr, P. & Zisserman, A. (1996), 3D model acquistions from extended image sequences, *in* '4th ECCV', 683–695.

Beardsley, P., Zisserman, A. & Murray, D. (1994), Navigation using affine structure from motion, *in* '3th ECCV', 85–96.

Bigün, J., Granlund, G. H. & Wiklund, J. (1991), 'Multidimensional orientation estimation with applications to texture analysis and optical flow', *PAMI* **13**(8), 775–790.

Bretzner, L. & Lindeberg, T. (1996), Feature tracking with automatic selection of spatial scales, ISRN KTH/NA/P--96/21--SE, KTH, Stockholm, Sweden.

Bretzner, L. & Lindeberg, T. (1997), On the handling of spatial and temporal scales in feature tracking, *in* 'Proc. 1st Scale-Space'97', Utrecht, Netherlands, 128–139.

Bretzner, L. & Lindeberg, T. (1998), Use your hand as a 3-d mouse, or, relative orientation from extended sequences of sparse point and line correspondences using the affine trifocal tensor. Technical report to be published at KTH.

Faugeras, O. (1992), What can be seen in three dimensions with a stereo rig?, *in* '2nd ECCV', 563–578.

Faugeras, O. (1995), 'Stratification of three-dimensional vision: Projective, affine and metric reconstructions', *JOSA* **12**(3), 465–484.

Faugeras, O. & Mourrain, B. (1995), On the geometry and algebra of the point and line correspondences between N images, *in* '5th ICCV', Cambridge, MA, 951–956.

Förstner, W. A. & Gülch, E. (1987), A fast operator for detection and precise location of distinct points, corners and centers of circular features, *in* 'ISPRS'.

Hartley, R. (1995), A linear method for reconstruction from points and lines, *in* '5th ICCV', Cambridge, MA, 882–887.

Heap, T. & Hogg, D. (1996), Towards 3D hand tracking using a deformable model, *in* 'Int. Conf. Autom. Face and Gesture Recogn., Killington, Vermont, 140–145.

Heyden, A. (1995), Reconstruction from image sequences by means of relative depth, *in* '5th ICCV', Cambridge, MA, 57–66.

Heyden, A., Sparr, G. & Åström, K. (1997), Perception and action using multilinear forms, *in* 'Proc. AFPAC'97', Kiel, Germany, 54–65.

Huang, T. S. & Lee, C. H. (1989), 'Motion and structure from orthographic projection', *IEEE-PAMI* **11**(5), 536–540.

Huang, T. S. & Netravali, A. N. (1994), 'Motion and structure from feature correspondences: A review', *Proc. IEEE* **82**, 251–268.

Koenderink, J. J. (1984), 'The structure of images', *Biol. Cyb.* **50**, 363–370.

Koenderink, J. J. & van Doorn, A. J. (1991), 'Affine structure from motion', *JOSA* 377–385.

Lee, J. & Kunii, T. L. (1995), 'Model-based analysis of hand posture', *Computer Graphics and Applications* pp. 77–86.

Lindeberg, T. (1994), *Scale-Space Theory in Computer Vision*, Kluwer, Netherlands.

Lindeberg, T. (1996), Edge detection and ridge detection with automatic scale selection, *in* 'CVPR'96', 465–470.

Lindeberg, T. & Bretzner, L. (1998), Visuellt människa-maskin-gränssnitt för tredimensionell orientering. Patent application.

Longuet-Higgins, H. C. (1981), 'A computer algorithm for reconstructing a scene from two projections', *Nature* **293**, 133–135.

Maybank, S. (1992), *Theory of Reconstruction from Image Motion*, Springer-Verlag.

McLauchlan, P., Reid, I. & Murray, D. (1994), Recursive affine structure and motion from image sequences, *in* '3th ECCV', Vol. 800, 217–224.

Morita, T. & Kanade, T. (1997), 'A sequential factorization method for recovering shape and motion from image streams', *IEEE-PAMI* **19**(8), 858–867.

Mundy, J. L. & Zisserman, A., eds (1992), *Geometric Invariance in Computer Vision*, MIT Press.

Quan, L. & Kanade, T. (1997), 'Affine structure from line correspondences with uncalibrated affine cameras', *IEEE-PAMI* **19**(8), 834–845.

Shapiro, L. S. (1995), *Affine analysis of image sequences*, Cambridge University Press.

Shashua, A. (1995), 'Algebraic functions for recognition', *IEEE-PAMI* **17**(8), 779–789.

Shashua, A. (1997), Trilinear tensor: The fundamental construct of multiple-view geometry and its applications, *in* 'Proc. AFPAC'97', Kiel, Germany, 190–206.

Spetsakis, M. E. & Aloimonos, J. (1990), 'Structure from motion using line correspondences', *IJCV* **4**(3), 171–183.

Sturm, P. & Triggs, B. (1996), A factorization based algorithm for multi-image projective structure and motion, *in* '4th ECCV', Vol. 1064, 709–720.

Tomasi, C. & Kanade, T. (1992), 'Shape and motion from image streams under orthography: A factorization method', *IJCV* **9**(2), 137–154.

Torr, P. H. S. (1995), Motion Segmentation and Outlier Detection, PhD thesis, Univ. of Oxford.

Ullman, S. (1979), *The Interpretation of Visual Motion*, MIT Press.

Ullman, S. & Basri, R. (1991), 'Recognition by linear combinations of models', *IEEE-PAMI* **13**(10), 992–1006.

Weng, J., Huang, T. S. & Ahuja, N. (1992), 'Motion and structure from line correspondences: Closed form solution and uniqueness results', *IEEE-PAMI* **14**(3), 318–336.

Xu, G. & Zhang, Z., eds (1997), *Epipolar Geometry in Stereo, Motion and Object Recognition: A Unified Approach*, Kluwer, Netherlands.