A Canonical Framework for Sequences of Images

Anders Heyden, Kalle Åström
Dept of Mathematics, Lund University
Box 118, S-221 00 Lund, Sweden
email: andersp@maths.lth.se kalle@maths.lth.se

Abstract

This paper deals with the problem of analysing sequences of images of rigid point objects taken by uncalibrated cameras. It is assumed that the correspondences between the points in the different images are known. The paper introduces a new framework for this problem. Corresponding points in a sequence of n images are related to each other by a fixed n-linear form. This form is an object invariant property, closely linked to the motion of the camera relative to the fixed world. We first describe a reduced setting in which these multilinear forms are easier to understand and analyse. This new formulation of the multilinear forms is then extended to the calibrated case. This formulation makes apparent the connection between camera motion, camera matrices and multilinear forms and the similarities between the calibrated and uncalibrated cases. These new ideas are then used to derive simple linear methods for extracting camera motion from sequences of images.

1 Introduction

A central problem in scene analysis is the analysis of 3D-objects from 2D-images, obtained by projections. In this paper we will concentrate on the case of a sequence of images consisting of points, with known correspondences. The objective is to calculate the shape of the object using the shapes of the images and to calculate the camera matrices, which gives the camera movement. We will present a method where no camera calibration is needed; making it possible to reconstruct the object and the camera movement up to a projective transformation.

During the last couple of years it has become increasingly popular to use invariance in computer vision. Two main types of invariance are viewpoint invariance and object invariance. Viewpoint invariant properties are properties which are intrinsic to the viewed object and independent of camera position or motion. An example is the classical cross ratio of four collinear points. Such properties can be used to recognise or reconstruct objects without knowing the relative position of the camera to the object. Another type of invariancy is the object invariant properties, i.e. properties of images which only depend on camera motion and not on the viewed object. An example of such a property is the epipolar constraint. These properties have twofold uses

- The constraints can be used to find further image correspondences.
- The constraints give information about the camera motion.

This type of invariancy is currently under intense research by a number of groups, cf. [1, 5, 6, 7, 8, 11, 12, 15]. The strong development has been aided by a very fruitful e-mail distributed discussion group during the spring. As is done in the work cited above, this paper deals with the generalisation of the epipolar constraint to the problem of analysing sequences of images. A new formulation of the multilinear constraints is presented which simplifies the analysis. This is first done in an affinely reduced setting closely related to the idea of affine shape, cf. [14], and relative affine reconstruction cf. [12]. The formulation is then generalised to the calibrated, the traditional uncalibrated and to a projectively reduced setting. This projective reduction has previously been used in [13] to simplify the estimation of the fundamental matrix. In this article the same idea is generalised to image sequences. This new formulation makes apparent the connection between camera motion, camera matrices and multilinear forms and the similarities between the calibrated and uncalibrated cases. The ideas are then used to derive simple linear methods for extracting camera motion from sequences of images. We hope that this new formulation will increase the understanding of different approaches and clarify the similarities between uncalibrated and calibrated cameras.

2 The affinely reduced setting

Given a sequence of images our first aim is to analyse the constraints on image points and on camera motion. These constraints will help us to determine camera motion from image correspondences without explicitly calculating the shape of the object points.

We assume that the camera is described by the following standard model,

$$\begin{bmatrix} \lambda_i x \\ \lambda_i y \\ \lambda_i \end{bmatrix} = P_i \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \quad i = 1, \ldots, n \tag{1}$$

where x and y are the image coordinates, P_i are the projection matrices, X, Y and Z are the coordinates of the object and λ_i are scale factors (different for different points and different images). Given a sequence of images from this model our aim is to reconstruct the object, to calculate mutual invariants between the images and to obtain a canonic description of this situation.

2.1 One Image

Since the camera matrices P are unknown we can multiply (1) from the left by an arbitrary nonsingular 3×3 matrix, which corresponds to choosing different affine coordinate systems in the images. A reasonable choice, to exploit this degree of freedom, is to make the first three points in the images to have affine coordinates $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$, that is a standard affine basis. Furthermore we can multiply P_i in (1) from the right by an arbitrary nonsingular 4×4 matrix, A, if we multiply each coordinate vector $(X,Y,Z,1)$ with A^{-1}. This corresponds to choosing affine or homogeneous coordinates in the object. In this way we can also assume that the scale factors for the first four points are equal to 1. For reasons that soon will be clear we choose the 4×4 matrix such that the first three points in the objects have homogeneous coordinates $(1,0,0,0)$, $(0,1,0,0)$, $(0,0,1,0)$ and the focus of the first camera has homogeneous coordinates $(0,0,0,1)$. This is basically the same idea as the relative affine coordinates in [11]. Putting together the equations from (1) for the first four points gives

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

since the only possibility for P_1 is $P_1 = [I \,|\, 0]$ and the focus of the camera maps to $(0,0,0)$. This also explains the choice of object coordinates. Because the homogeneous coordinates are determined up to an unknown scale factor we can choose this scale factor such that all scale factors, λ_1 in (1) in the first image are equal. It also follows that an arbitrary point in the object with homogeneous coordinates (X,Y,Z,W) projects to the image point with affine coordinates (X,Y,Z). This means that the only unknown coordinates are the fourth and if these coordinates can be calculated we have a reconstruction.

2.2 Two Images

We now consider a second image. Let the scale factors, λ_2 in (1), for the first three points be p_1, p_2, p_3 and let p be the scale factor for an arbitrary point with affine coordinates (a_1,b_1,c_1) in the first image and (a_2,b_2,c_2) in the second image. Then the camera matrix is given from

$$\begin{bmatrix} p_1 & 0 & 0 & pa_2 \\ 0 & p_2 & 0 & pb_2 \\ 0 & 0 & p_3 & pc_2 \end{bmatrix} = \begin{bmatrix} p_1 & 0 & 0 & \mathbf{t}_1 \\ 0 & p_2 & 0 & \mathbf{t}_2 \\ 0 & 0 & p_3 & \mathbf{t}_3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & a_1 \\ 0 & 1 & 0 & b_1 \\ 0 & 0 & 1 & c_1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where $(\mathbf{t}_1,\mathbf{t}_2,\mathbf{t}_3) = (pa_2 - p_1a_1, pb_2 - p_2b_1, pc_2 - p_3c_1)$. That is $P_2 = [D_p \,|\, \mathbf{t}]$, where D_p is the diagonal matrix obtained from (p_1,p_2,p_3) and $\mathbf{t} = (\mathbf{t}_1,\mathbf{t}_2,\mathbf{t}_3)$.

Consider now an arbitrary point (X,Y,Z,W) in the object and its projection (x_1,y_1,z_1) and (x_2,y_2,z_2) in the first and second image. It follows from (2) and (3) that

$$\begin{aligned} (x_1,y_1,z_1) &= (X,Y,Z), \\ (px_2,py_2,pz_2) &= (p_1X + \mathbf{t}_1W, p_2Y + \mathbf{t}_2W, p_3Z + \mathbf{t}_3W), \end{aligned} \tag{4}$$

where p is the scale factor for the point. From these equation we can eliminate p and (X,Y,Z,W) which gives

$$\det \begin{bmatrix} p_1x_1 & \mathbf{t}_1 & x_2 \\ p_2y_1 & \mathbf{t}_2 & y_2 \\ p_3z_1 & \mathbf{t}_3 & z_2 \end{bmatrix} = 0. \tag{5}$$

Rewriting this determinant gives

$$\begin{bmatrix} x_1 & y_1 & z_1 \end{bmatrix} \begin{bmatrix} 0 & p_1\mathbf{t}_3 & -p_1\mathbf{t}_2 \\ -p_2\mathbf{t}_3 & 0 & p_2\mathbf{t}_1 \\ p_3\mathbf{t}_2 & -p_3\mathbf{t}_1 & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix} = 0, \tag{6}$$

or $x^T F \bar{x} = 0$, where x and $\bar{x}$ are the affine coordinate vectors in image 1 and image 2 and F are the fundamental matrix for this particular choice of coordinates. We call F the **reduced fundamental matrix** and $\mathbf{t}$ the **projective translational vector**. It can be seen from (6) that F can be factorised as

$$F = \begin{bmatrix} p_1 & 0 & 0 \\ 0 & p_2 & 0 \\ 0 & 0 & p_3 \end{bmatrix} \begin{bmatrix} 0 & \mathbf{t}_3 & -\mathbf{t}_2 \\ -\mathbf{t}_3 & 0 & \mathbf{t}_1 \\ \mathbf{t}_2 & -\mathbf{t}_1 & 0 \end{bmatrix} \tag{7}$$

that is as a product of a diagonal and a skew matrix. The diagonal matrix contains the kinetic depths, see [14] and [7], on the diagonal and the skew matrix is formed from the vector $\mathbf{t}$. From (6) and (7) it is obvious that $\mathbf{t} = \mu e_{21}$ where e_{21} are the affine coordinates of the epipole in the second image and $D_p^{-1}\mathbf{t} = \nu e_{12}$ where e_{12} are the affine coordinates of the epipole in the first image, and μ and ν are scale factors needed to make the sum of the affine coordinates equal to 1. The components of the reduced fundamental matrix can be calculated linearly from eight point matches and then the kinetic depths and the epipoles can be recovered linearly. Finally, the eliminated W-coordinates can be calculated, which gives the reconstruction.

2.3 Three Images

We now consider three images concurrently. Let the scale factors, λ_3 in (1), for the first three points in the third image be q_1, q_2, q_3 and the affine coordinates for a fourth arbitrary point in the third image be (a_3, b_3, c_3), with scale factor q. Then the camera matrix for camera three is given by

$$P_3 = \begin{bmatrix} q_1 & 0 & 0 & \bar{\mathbf{t}}_1 \\ 0 & q_2 & 0 & \bar{\mathbf{t}}_2 \\ 0 & 0 & q_3 & \bar{\mathbf{t}}_3 \end{bmatrix} \tag{8}$$

where $(\bar{\mathbf{t}}_1, \bar{\mathbf{t}}_2, \bar{\mathbf{t}}_3) = (qa_3 - q_1a_1, qb_3 - q_2b_1, qc_3 - q_3c_1)$. It follows that $P_3 = [D_q \,|\, \bar{\mathbf{t}}]$, where D_q is the diagonal matrix obtained from (q_1, q_2, q_3) and $\bar{\mathbf{t}} = (\bar{\mathbf{t}}_1, \bar{\mathbf{t}}_2, \bar{\mathbf{t}}_3)$ is the projective translational vector between image 1 and image 3. Consider now an arbitrary point (X, Y, Z, W) in the object and its projection (x_3, y_3, z_3) in the third image. It follows from (3) that

$$(qx_3, qy_3, qz_3) = (q_1X + \bar{\mathbf{t}}_1W, q_2Y + \bar{\mathbf{t}}_2W, q_3Z + \bar{\mathbf{t}}_3W), \tag{9}$$

where q is the scale factor for the point. From (4) and (9) we can eliminate p, q and (X, Y, Z, W) which gives

$$\operatorname{rank} \begin{bmatrix} p_1x_1 & \mathbf{t}_1 & x_2 & 0 \\ p_2y_1 & \mathbf{t}_2 & y_2 & 0 \\ p_3z_1 & \mathbf{t}_3 & z_2 & 0 \\ q_1x_1 & \bar{\mathbf{t}}_1 & 0 & x_3 \\ q_2y_1 & \bar{\mathbf{t}}_2 & 0 & y_3 \\ q_3z_1 & \bar{\mathbf{t}}_3 & 0 & z_3 \end{bmatrix} = 3. \tag{10}$$

From this equation it follows that there are three algebraically independent equations in the image coordinates, see also [11]. Two of them can be chosen as the bilinear constraints from the reduced fundamental matrices and one is a combined expression involving coordinates of points in all three images. It can also be shown that there are four linearly independent trilinear expressions in the image coordinates. It follows from (10) that it is possible to recover p, q, $\mathbf{t}$, and $\bar{\mathbf{t}}$ up to scale factors. If we fix the scale for p, q and $\mathbf{t}$ then $\bar{\mathbf{t}}$ can be recovered without unknown scale factor. This means that the camera matrices can be completely recovered from this information. We remark that this can be done linearly from at least seven point matches in three images by the reduced fundamental tensor, cf. [7].

Considering the pair of image 2 and image 3, we can derive the reduced fundamental matrix from (10) by eliminating x_1, x_2 and x_3 which gives

$$\begin{aligned} &\det \begin{bmatrix} q_1x_2 & p_1x_3 & q_1\mathbf{t}_1 - p_1\bar{\mathbf{t}}_1 \\ q_2y_2 & p_2y_3 & q_2\mathbf{t}_2 - p_2\bar{\mathbf{t}}_2 \\ q_3z_2 & p_3z_3 & q_3\mathbf{t}_3 - p_3\bar{\mathbf{t}}_3 \end{bmatrix} = \\ &= p_1p_2p_3 \det \begin{bmatrix} r_1x_2 & x_3 & r_1\mathbf{t}_1 - \bar{\mathbf{t}}_1 \\ r_2y_2 & y_3 & r_2\mathbf{t}_2 - \bar{\mathbf{t}}_2 \\ r_3z_2 & z_3 & r_3\mathbf{t}_3 - \bar{\mathbf{t}}_3 \end{bmatrix} = 0, \end{aligned} \tag{11}$$

where $r_i = q_i/p_i$ are the kinetic depths between image 2 and image 3. The epipole of camera 2 in image 3 is a multiple of $\hat{\mathbf{t}} = (r_1\mathbf{t}_1 - \bar{\mathbf{t}}_1, r_2\mathbf{t}_2 - \bar{\mathbf{t}}_2, r_3\mathbf{t}_3 - \bar{\mathbf{t}}_3)$, the projective translational vector between image 2 and image 3. This means that if $\mathbf{t}$, $\bar{\mathbf{t}}$ and $\hat{\mathbf{t}} = D_r\mathbf{t} - \bar{\mathbf{t}}$ can be determined up to unknown scale factors from reduced fundamental matrices between the three different pairs of images, then it is possible to recover the camera matrices if $\mathbf{t}$, $\bar{\mathbf{t}}$ and $\hat{\mathbf{t}}$ are not collinear. Geometrically this can be viewed as vector addition of $D_r\mathbf{t}$ and $\hat{\mathbf{t}}$ with appropriate lengths giving $\bar{\mathbf{t}}$ also with appropriate length.

2.4 Four Images

In the same way as above it can be shown that considering four images, elimination of object coordinates and kinetic depths gives

$$\operatorname{rank} \begin{bmatrix} p_1x_1 & \mathbf{t}_1 & x_2 & 0 & 0 \\ p_2y_1 & \mathbf{t}_2 & y_2 & 0 & 0 \\ p_3z_1 & \mathbf{t}_3 & z_2 & 0 & 0 \\ q_1x_1 & \bar{\mathbf{t}}_1 & 0 & x_3 & 0 \\ q_2y_1 & \bar{\mathbf{t}}_2 & 0 & y_3 & 0 \\ q_3z_1 & \bar{\mathbf{t}}_3 & 0 & z_3 & 0 \\ s_1x_1 & \tilde{\mathbf{t}}_1 & 0 & 0 & x_4 \\ s_2y_1 & \tilde{\mathbf{t}}_2 & 0 & 0 & y_4 \\ s_3z_1 & \tilde{\mathbf{t}}_3 & 0 & 0 & z_4 \end{bmatrix} = 4, \tag{12}$$

where s_i are the kinetic depths between image 1 and 4, $\tilde{\mathbf{t}}_i$ is the projective translational vector between image 1 and 4 and x_4, y_4 and z_4 are affine coordinates in the fourth image. By considering all five by five subdeterminants it can be seen that we get nothing new, just bilinearities from pairs of images and trilinearities from triples of images.

2.5 Sequence of Images

It has been shown above that the camera matrices can be written as

$$P_1 = [I \,|\, 0], \qquad P_i = [D_i \,|\, -\mathbf{t}_i], \quad i = 2 \ldots n, \tag{13}$$

where D_i are diagonal matrices formed from the kinetic depths from the three basis points, $\mathbf{t}_i$ are multiples of the affine coordinates for the epipoles of camera 1 in image i and $n + 1$ is the number of images. Equivalently this can be written as

$$P_1 = [I \,|\, 0], \qquad P_i = [D_i \,|\, -D_i\bar{\mathbf{t}}_i], \quad i = 2 \ldots n, \tag{14}$$

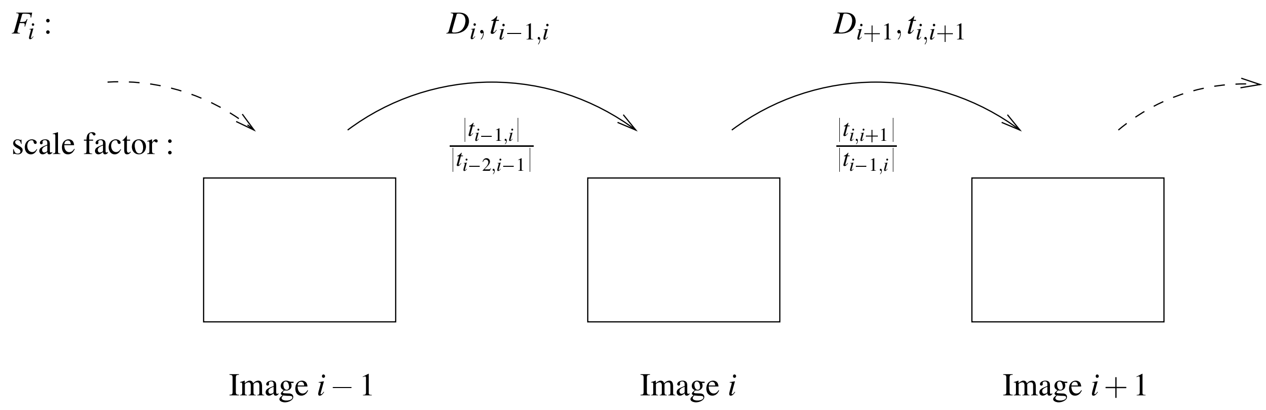

Figure 1: The reduced fundamental matrices and the extra scale factors needed to uniquely describe an image sequence.

where $\bar{\mathbf{t}}_i$ is a multiple of the affine coordinates of the epipole in image 1 from camera i. Notice the similarities with the calibrated case, see also [1]. Now it is possible to calculate the positions of the center, Z_i, of camera i as the nullspace of P_i, which is $(\bar{\mathbf{t}}_i, 1) = (\bar{\mathbf{t}}_{i1}, \bar{\mathbf{t}}_{i2}, \bar{\mathbf{t}}_{i3}, 1)$, where the projective translational vector $\bar{\mathbf{t}}_i$ is a multiple of the affine coordinates of the epipole in image 1 of camera i. Then all possible camera locations can be calculated from this representation by a projective transformation.

It has also been shown that the reduced fundamental matrices between consecutive pairs of images contains nearly all information. The extra information needed is just a scale factor, the ratio between the length of the projective translational vectors. These ratios can be obtained either from the trilinearities or from other reduced fundamental matrices. This can be visualised as in Figure 1, cf also [8].

3 Projectively reduced setting

If four or more coplanar points can be seen and detected in each image, then the image of these can be used to simplify the problem in a similar way as in the discussion above. Choose a projective coordinate system so that the coplanar points lie on the plane at infinity and have coordinates $(1,0,0,0)$, $(0,1,0,0)$, $(0,0,1,0)$ and $(1,1,1,0)$. Then choose a projective coordinate system in each image so that the image of the four points have coordinates $(1,0,0)$, $(0,1,0)$, $(0,0,1)$ and $(1,1,1)$. Then each camera matrix must be of the form

$$P_i = \begin{bmatrix} 1 & 0 & 0 & \mathbf{t}_1^x \\ 0 & 1 & 0 & \mathbf{t}_1^y \\ 0 & 0 & 1 & \mathbf{t}_1^z \end{bmatrix} . \qquad (15)$$

This can be verified by a explicit calculation or by noting that the projection from the plane at infinity to the image plane is governed by the first 3×3 block Q in the camera matrix. Since this transformation is the identity on a projective basis, the matrix Q must be the identity. The last 3×1 block is undetermined. Thus we have reduced the effect of camera rotation and change in internal calibration. After reduction the image sequence can be analysed as if the cameras motion was purely translational. The same effect is obtained as soon as the planar object is aligned in each image. The specific choice of coordinate system, e.g. $(1,0,0)$, $(0,1,0)$, $(0,0,1)$ and $(1,1,1)$, is not important. The multilinear constraints now become of the form

$$\operatorname{rank} \begin{bmatrix} x_1 & \mathbf{t}_{12} & x_2 & \dots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 \\ x_1 & \mathbf{t}_{1N} & 0 & \dots & x_N \end{bmatrix} = N. \qquad (16)$$

This is linear in the motion and can thus be solved by linear methods. In this way the problem of determining the camera direction and motion is split into two independent parts. The direction is estimated using the points at infinity and the position is estimated using the remaining points.

4 Generalisations of the above

The analysis above can now easily be summarised or reformulated in four settings.

1. The traditional uncalibrated setting.
2. The affinely reduced uncalibrated setting.
3. The projectively reduced uncalibrated setting.
4. The internally calibrated setting.

The affinely reduced setting is the one described in Section 2, where the image of three corresponding points are used as an affine basis in each image. The projectively reduced setting is the setting described in section 3, where four or more coplanar points are used as a projective basis in each image plane.

The multilinear constraints can be written as

$$\operatorname{rank}\begin{bmatrix} x_1 & \mathbf{t}_{12} & Q_{12}x_2 & \dots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 \\ x_1 & \mathbf{t}_{1N} & 0 & \dots & Q_{1N}x_N \end{bmatrix} = N, \qquad (17)$$

where $x_1, \dots, x_N$ are the corresponding image of a point in N images, and Q_{ij} are 3×3 matrices. These matrices can be thought of as the image coordinate changes from image j to image i, or the relative change of direction between time instant i and j. The only difference between the four settings lie in the choice of manifold for Q_{ij}. These matrices are:

1. General planar projective transformations in setting 1. (8 degrees of freedom).
2. Diagonal matrices in setting 2. (2 d o f).
3. Identity matrices in setting 3. (0 d o f).
4. Rotation matrices in setting 4. (3 d o f).

Notice that the form (17) is a summary of all multilinear constraints on a sequence of N images. Each $N-1 \times N-1$ subdeterminant is a multilinear form in the image coordinates. Since the rank of the whole matrix should be zero, each such subdeterminant should be zero. The constraints can be used in several ways. First of all the parameters $t_{12}, \dots, t_{1N}$ and $Q_{12}, \dots, Q_{1N}$ can be calculated given enough image correspondences. In setting 1 this can only be done uniquely up to a choice of the plane at infinity. However, in settings 2, 3 and 4 the plane at infinite is already well defined so the parameters $\mathbf{t}_{12}, \dots, \mathbf{t}_{1N}$ and $Q_{12}, \dots, Q_{1N}$ can be found essentially uniquely. Second, knowledge of the parameters makes it possible to find further image correspondences. Third, the parameters are closely linked to a reconstruction of the camera matrices, $P_i = Q_{1i}[I \mid -\mathbf{t}_{1i}]$. Notice that the projection onto the screen can be seen as a composition of a standard projection towards the focal point t_{1i} and a projective rearrangement Q_{1i} of the image coordinates. Depending on the setting this projective rearrangement have 0, 2, 3 or 8 degrees of freedom. Another interpretation is that t_{1i} represents the position of the camera and Q_{1i} the generalised orientation of the camera.

As in the previous discussion it is easy to see that all non-trivial $N+1 \times N+1$ subdeterminant is a product of either a bilinear or a trilinear constraint. Thus the trilinear constraints

$$\operatorname{rank}\begin{bmatrix} u_1 & \mathbf{t}_{12} & Q_{12}u_2 & 0 \\ u_1 & \mathbf{t}_{13} & 0 & Q_{13}u_3 \end{bmatrix} = 3$$

and the bilinear constraints

$$\operatorname{rank}\begin{bmatrix} u_1 & \mathbf{t}_{12} & Q_{12}u_2 \end{bmatrix} = 2$$

are the building blocks on which the motion can be calculated. As usual the bilinear constraints can be represented with the fundamental matrix,

$$u_1^T F_{12} u_2 = u_1^T T_{\mathbf{t}_{12}} Q_{12} u_2 = 0,$$

where $T_{\mathbf{t}}$ is the matrix representing cross product with vector $\mathbf{t}$.

Notice the strong similarity between the four settings. The only difference is a priori knowledge on the matrices Q_{ij}.

Denote by M_i the coordinate change from world coordinates to the i'th camera coordinates, cf. Fig. 2,

$$M_i = \begin{pmatrix} Q_i & -Q_i\mathbf{t}_i \\ 0 & 1 \end{pmatrix} \qquad (18)$$

with this notation each camera can be considered as a coordinate transformation followed by the standard camera matrix

$$P_0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \qquad (19)$$

so that $P_i = P_0 M_i$.

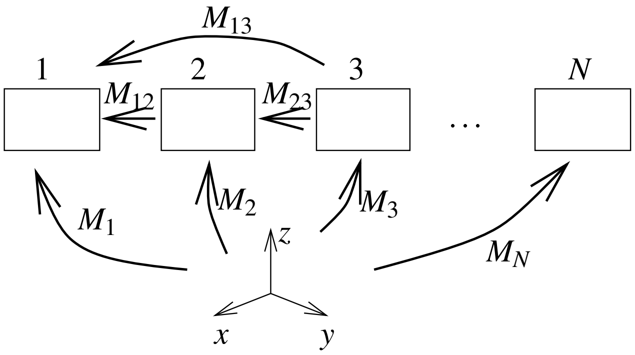

Figure 2: A sequence of images and coordinate transformations between different camera coordinate systems and the object coordinate system.

The situation is illustrated in Fig. 2. In this figure we have also included the change of coordinate system between different camera coordinates M_{ij} which we define as $M_{ij} = M_i M_j^{-1}$. This block matrix can be seen to be

$$M_{ij} = M_i M_j^{-1} = \begin{pmatrix} Q_{ij} & \mathbf{t}_{ij} \\ 0 & 1 \end{pmatrix}, \qquad (20)$$

where

$$Q_{ij} = Q_i Q_j^{-1} \quad \text{and} \quad \mathbf{t}_{ij} = Q_i(\mathbf{t}_j - \mathbf{t}_i). \qquad (21)$$

Thus Q_{ij} is the change of 'orientation' from the j'th camera coordinate system to the i'th camera coordinate system. Through the multilinear forms, which can be calculated from the images, we obtain partial information about the relative coordinate transformations M_{ij}. This partial information is often difficult to use and analyse. It is therefore natural to try to estimate the full motion of the cameras instead, i.e. the coordinate transformations M_i, $i = 1, \ldots, n$.

The bilinear forms can be calculated from point and/or curve correspondences in the two images, cf. [3]. Thus $(Q_{ij}, \mathbf{t}_{ij})$ can be calculated, but $\mathbf{t}_{ij}$ can only be calculated up to an unknown scale factor and in the totally uncalibrated setting Q_{ij} can only be calculated up to a transformation related to the choice of plane at infinity. This information can then be used both to find further image correspondences but also as we shall see to calculate the full camera motion, up to an unknown scale factor. A collection of bilinear constraints are consistent if and only if

$$\begin{cases} Q_{ij} = Q_{ik} Q_{kj} \\ \mathbf{t}_{ij}, \mathbf{t}_{ik}, Q_{ik}\mathbf{t}_{kj} \text{ are coplanar} \end{cases} \qquad (22)$$

If the cameras are in general position the fact that the three directions $\mathbf{t}_{ij}$, $\mathbf{t}_{ik}$, $Q_{ik}\mathbf{t}_{kj}$ are coplanar but not collinear can be used to reconstruct the camera motion without using the trilinear constraints. If on the other hand $\mathbf{t}_{ij}, \mathbf{t}_{ik}, Q_{ik}\mathbf{t}_{kj}$ are collinear the trilinear constraints are needed for the reconstruction.

5 Structure and motion from image sequences

One of the main motives for an investigation of the multilinear constraints, as is done above, is to find new solutions to the problem of extracting structure and motion from image sequences. Given a number of images of an unknown scene the best estimate of the structure and motion can be obtained by optimising over all parameters as in [4] and [2]. This is necessary if a very good reconstruction is needed as in surveying and the kind of application described in [2]. But in other applications a recursive approach is more appropriate. This is advantageous for two reasons: (i) the number of parameters to estimate does not grow all the time, (ii) the amount of processed data does not grow all the time. One idea would be to recursively update the motion (the camera matrix) and the structure (the reconstruction) at each time instant as is done in [9]. This is still a very large amount of data to update. Another idea is to recursively update the motion (the camera matrix) only. This is where the study of multilinear constraints are needed. The discussion above shows that the trilinear forms of images $k-2, k-1$ and k are needed to update camera matrix P_k. So a discrete time filter would have to be of type

$$(P_{k-3}, P_{k-2}, P_{k-1}, T_{k-2,k-1,k}) \rightarrow (P_{k-2}, P_{k-1}, P_k),$$

where $T_{k-2,k-1,k}$ represents the information in the trilinear constraints. The investigation above also indicate that in the continuous time case a second order filter is needed, i.e. second derivatives with respect to time are needed to estimate motion.

Another area where the discussion above is needed is to understand how motion can be estimated from multilinear constraints when these have been found without point correspondences, as in [3].

6 Experiments

Some preliminary investigations were performed to evaluate the ideas above. The first thing to note is that the investigation above only deals with the ideal noise-free case. While experimenting noise have to be taken into account. In these experiment linear methods have been used combined with the singular value decomposition. This gives fast algorithms and reasonable tolerance to noise. The solutions from these fast but not so robust linear algorithms can then be used as initial estimates to better, but slower non-linear estimators.

The affinely reduced setting.

We illustrate the affinely reduced setting with an experiment where 9 images have been taken of a simple scene. These are shown in Fig. 3. Some of the corners are extracted and three corners, one from each box on the floor, are used as an affine basis in each image. The paths that each corner makes can now be analysed in the affinely reduced setting. Linear methods are then used to estimate the kinetic depths and the camera motion. This motion is presented together with the reconstruction of some of the corner points.

The projectively reduced setting.

To illustrate the idea of projective reduction, we have taken a series of images with a planar curve and a few simple objects. Five of these images are shown in Fig. 4. The planar curve and some of the corners are extracted and superimposed in the same figure. Each corner makes a path in the image plane. These paths fulfill the multilinear constraints in setting 1. To estimate motion both directions matrices Q_{ij} and camera positions t_{ij} has to be estimated. The planar object is detected and a projective coordinate system is chosen at each time instant so that the planar curve has the same coordinates. The paths that each corner makes can now be analysed as if they were taken by a purely translating camera, cf. Eq. 16. Thus simple linear methods can be used to calculate camera motion, t_{1j}. When the full camera motion is known it is easy to reconstruct the 3D point configuration. The camera motion is shown together with a reconstruction of the curve and the corners.

The affinely and projectively reduced setting share a common numerical disadvantage. The simplification is achieved by dividing the problem into two parts. The basis points are used to estimate the direction Q_{ij} and the remaining features are used to estimate the position t_{ij}. It would be numerically better to use all available measurements to solve both direction and position. Still the reduced setting is important for theoretical reasons. We believe that a clearer understanding of the problem is obtained through this formulation. The connections between affine shape, cf. [14], relative affine reconstruction, cf. [12] and the traditional uncalibrated and calibrated setting is more apparent. The reduced setting can also be useful to get initial estimates of motion and reconstruction. Refinements can then be made using the whole material in an appropriate way.

7 Conclusions and Acknowledgement

In this paper is presented, simplifications and a new formulation of the multilinear forms that exist between a sequence of images. It becomes apparent that multilinear forms contain information in the bilinear and trilinear forms only. Furthermore the bilinear forms are in general sufficient to use and the trilinear forms often follow from the bilinear forms. This representation is fairly close to the representation of the motion and it is easy to generalise to different settings. Four such settings calibrated, uncalibrated, affinely reduced and projectively reduced, are described in the paper.

The affinely reduced setting is described in detail. Much is simplified through the choice of affine basis in each image. A new form of the fundamental matrix, called the reduced fundamental matrix, has been given. It has been shown that it can be factorised as a product of a diagonal and a skew matrix.

Further simplifications can be achieved if four or more coplanar points can be identified in a short subsequence. These coplanar points can be used in a projectively reduced setting. After the reduction the analysis of the remaining points can be made as if the camera underwent a purely translating motion, yielding even simpler forms on the fundamental matrices and the trilinearities.

Two short image sequences are analysed with the methods described in the paper. Camera movement as well as the reconstruction of the viewed object are estimated using the two reduced settings, thus demonstrating the feasibility of the linear approach on real image data.

Further work in this area would be to incorporate stochastic analysis and proper camera motion estimators using this new framework.

The work has been done within the ESPRIT-BRA Viewpoint Invariant Visual Acquisition (VIVA).

References

[1] Åström. K., Heyden, A., Canonic Framework for Sequences of Images: Similarities between Calibrated and Uncalibrated Case, *Proc. Symposium on Image Analysis, Linköping, Sweden, 1995.*

[2] Åström. K., Automatic Mapmaking, *Proc. First IFAC International Conference on Intelligent Autonomous Vehicles, Southampton, UK 1993.*

[3] Cipolla, R., Åström. K., Giblin, P., Motion from the frontier of curved surfaces, *Proc. 5'th ICCV, 1995.*

[4] Faugeras, O., D., What can be seen in three dimensions with an uncalibrated stereo rig?, *ECCV'92, Lecture notes in Computer Science, Vol 588. Ed. G. Sandini, Springer-Verlag* 1992, pp. 563-578.

[5] Faugeras, O., Mourrain, B., On the geometry and algebra of the point and line correspondences between N images, *Proc. 5'th ICCV, 1995.*

[6] Heyden, A., Reconstruction from Three Images of Six Point Objects, *Proc. Symposium on Image Analysis, SSAB, Halmstad, Sweden, 1994.*

[7] Heyden, A., Reconstruction from Image Sequences by means of Relative Depths, *Proc. 5'th ICCV, 1995.*

[8] Luong, Q., Vieville, T., Canonic Representations for the Geometries of Multiple Projective Views, *ECCV'94, Lecture notes in Computer Science, Vol 800. Ed. Jan-Olof Eklund, Springer-Verlag* 1994, pp. 589-599.

[9] McLauchlan, P. F., Murray, D. W., A unifying framework for structure and motion recovery from image sequences, *Proc. 5'th ICCV, 1995.*

[10] Quan, L., Invariants of 6 points from 3 uncalibrated images, *ECCV'94, Lecture notes in Computer Science, Vol 801. Ed. Jan-Olof Eklund, Springer-Verlag* 1994, pp. 459-470.

[11] Shashua, A., Trilinearity in Visual Recognition by Alignment, *ECCV'94, Lecture notes in Computer Science, Vol 800. Ed. Jan-Olof Eklund, Springer-Verlag* 1994, pp. 479-484.

[12] Shashua, A., Navab, N., Relative Affine Structure: Theory and Application to 3D Reconstruction from Perspective Views, *CVPR94*, 1994, pp. 483-489.

[13] Sinclair, D., Christensen, H., Rothwell, C., Using the relationship between a plane projectivity and the fundamental matrix, *9'th SCIA, Uppsala, Sweden*, 1994, pp. 181-188.

[14] Sparr, G., A Common Framework for Kinetic Depth, Reconstruction and Motion for Deformable Objects, *ECCV'94, Lecture notes in Computer Science, Vol 801. Ed. J-O. Eklund, Springer-Verlag* 1994, pp. 471-482.

[15] Triggs, B., Matching Constraints and the Joint Image, *Proc. 5'th ICCV*, 1995.

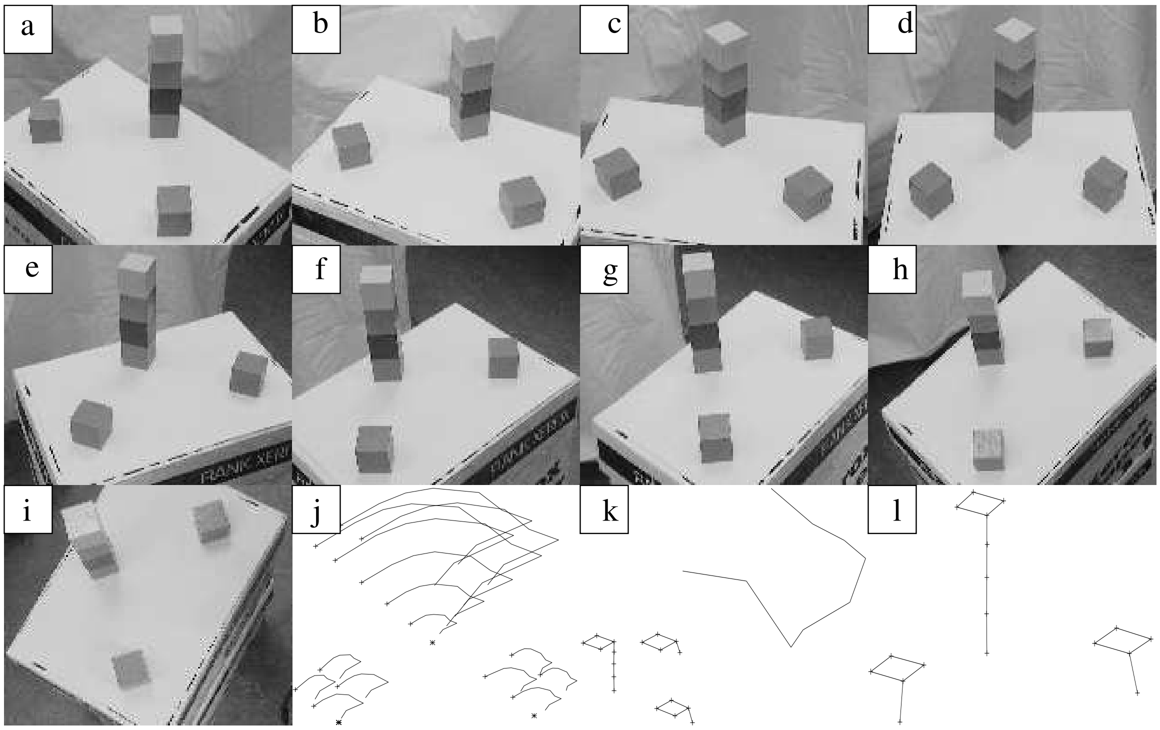

Figure 3: Illustration of the affinely reduced setting. Fig. 3.a-i show nine images of a simple scene. In Fig. 3.j, some of the extracted points are shown in the affine coordinate system defined by three basis points. In Fig. 3.k, the reconstructed camera motion is shown together with a reconstruction of the extracted corner points. Fig. 3.l highlights the reconstructed object.

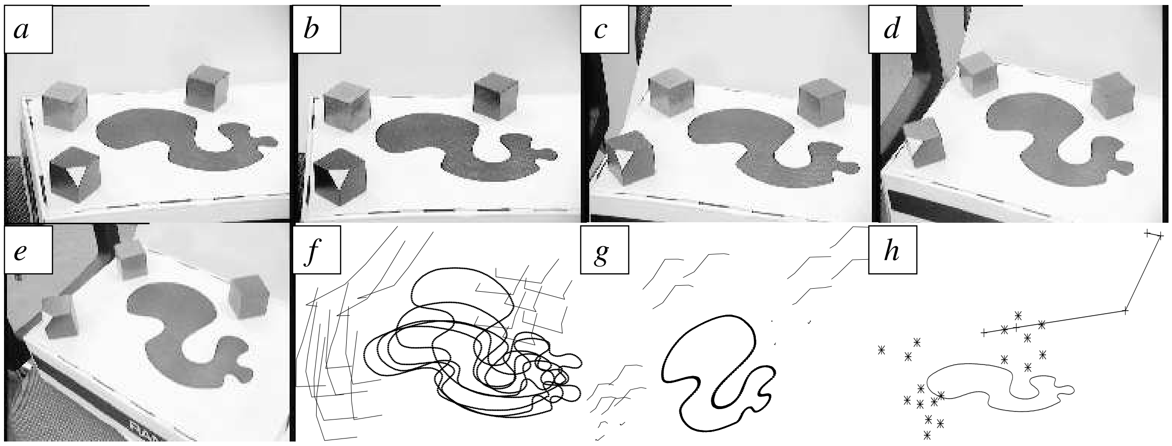

Figure 4: Illustration of the projectively reduced setting. Fig. 4.a-e show five images of a simple scene. In Fig. 4.f the planar curve and some of the corners are extracted and superimposed in the same figure. In Fig. 4.g the planar object is used as a basis for projective coordinate system in each image. The point paths can now be analysed as if they were taken by a purely translating camera. Fig. 4.h shows the reconstructed camera motion together with a reconstruction of the curve and the corners.