ESTIMATION OF AND VIEW SYNTHESIS WITH THE TRIFOCAL TENSOR

Helmut Mayer
Institute for Photogrammetry and Cartography, Bundeswehr University Munich, D-85577 Neubiberg, Germany
Helmut.Mayer@UniBw-Muenchen.de

KEY WORDS: Visualization, Orientation, Hierarchical, Matching, Geometry

ABSTRACT

In this paper we propose a robust hierarchical approach for the estimation of the trifocal tensor. It makes use of pyramids, the sub-pixel Förstner point operator, least squares matching, RANSAC, and the Carlsson-Weinshall duality. We also show how the trifocal tensor can be utilized for an efficient view synthesis which we have optimized by parameterizing it according to the epipolar lines.

1 INTRODUCTION

In photogrammetry relative orientation and bundle block adjustment employing the collinearity equation are standard procedures. The reasons why additional orientation procedures based on projective geometry nevertheless are a viable alternative and partly are even necessary are:

- These orientation procedures make direct solutions available, i.e., no approximate values and no iterations are needed. This is often essential for automatic procedures in variable geometry close range applications.
- With these orientation procedures linear relations between measurements in the images arise without the need to reconstruct three dimensional (3D) geometry explicitly. This is especially important for applications such as video communication, where the goal is to generate artificial views and not 3D geometry. Representation and projection based on projective geometry are standard in computer graphics.

Linear projective relations are known in photogrammetry for quite a while theoretically, though they have seldom been used in practice. An account of the history of linear methods in photogrammetry is, e.g., given in (Brandstätter, 1996).

In this paper we concentrate on two things: The robust estimation of the trifocal tensor from three images and its utilization for view synthesis. Our major achievements are as follows:

- By means of a hierarchical approach based on image pyramids we reduce the search space. Efficiency but also robustness are improved considerably. Highly precise conjugate points are obtained from a least-squares matching of points obtained from the Förstner operator (Förstner and Gülch, 1987)
- We use the Carlsson-Weinshall duality to calculate a solution for the trifocal tensor from a minimum of six point triplets. This is the basis for a RANSAC (Fischler and Bolles, 1981) based robust algorithm.
- We have optimized the view synthesis scheme proposed in (Avidan and Shashua, 1998) by linearly projecting the points as proposed in (Hartley and Zisserman, 2000) and, particularly, by parameterizing the points according to the epipolar lines.

In Section 2 we introduce basic concepts and notations. They comprise the fundamental matrix as it is used as a basic building block of our orientation procedure and the essential matrix as the view synthesis relies on calibration. Section 3 describes the trifocal tensor and in Section 4 the point transfer based on the trifocal tensor is detailed. In Section 5 we show how the trifocal tensor can be estimated from image data. Section 6 summarizes the algorithm we use to estimate depth, i.e., disparity, from two images. This is the basis for view synthesis presented in Section 7. We end up with conclusions.

2 BASICS OF LINEAR ORIENTATION

2.1 Homogeneous Coordinates

Homogeneous coordinates are derived from Euclidean coordinates by adding an additional coordinate and free scaling. Generally, for two dimensional (2D) and 3D points holds (Hartley and Zisserman, 2000):

$$\mathbf{x} = \lambda \begin{pmatrix} \boldsymbol{x} \\ 1 \end{pmatrix}, \quad \lambda \neq 0 \quad .$$

In our notation we distinguish homogeneous 2D and 3D vectors $\mathbf{x}$ and $\mathbf{X}$, respectively, as well as matrices $\mathsf{\mathbf{P}}$, which represent the same object also after a change of the scaling factor λ (bold), from Euclidean vectors $\boldsymbol{x}$ and $\boldsymbol{X}$ as well as matrices $\boldsymbol{R}$ (bold italics).

Straight lines $\mathbf{l}$ in 2D and planes $\mathbf{P}$ in 3D can also be described by parameters termed homogeneous coordinates. The incidence relation $ax + by + c = 0$ of point $\mathbf{x}$ and straight line $\mathbf{l}$, i.e., the point $\mathbf{x}$ lies on the straight line $\mathbf{l}$, reads with homogenous parameters for the straight line $\mathbf{l} = (abc)^\mathsf{T}$ $\quad \mathbf{x}^\mathsf{T}\mathbf{l} = \mathbf{l}^\mathsf{T}\mathbf{x} = \mathbf{x} \cdot \mathbf{l} = 0$.

2.2 Perspective Transformation

In a local spatial coordinate system with origin in the camera (projection) center O and with the image plane given by the equation $z = c$, the 3D object point $P(X, Y, Z)$ is projected to the image point $P'(x', y')$ by $x' = cX/Z + x'_h$, $y' = cY/Z + y'_h$.

For the general case holds equation (1) given the point in object space $\mathbf{X}$. The exterior orientation is described by projection center $O(\mathbf{X}_0)$ and rotation matrix $\boldsymbol{R}$. The interior orientation is modeled by principal distance c, principal point (x'_h, y'_h), scale difference m of the coordinate axes and skew of the axes s.

The 5 parameters chosen for the interior orientation are collected into the calibration matrix:

$$\boldsymbol{K} = \begin{pmatrix} c & cs & x'_h \\ 0 & c(1+m) & y'_h \\ 0 & 0 & 1 \end{pmatrix} \quad .$$

$$\mathbf{x}' = \begin{pmatrix} 1 & s & x'_h \\ 0 & 1+m & y'_h \\ 0 & 0 & 1 \end{pmatrix} \left(\begin{array}{ccc|c} c & 0 & 0 & 0 \\ 0 & c & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right) \begin{pmatrix} \boldsymbol{R} & \mathbf{0} \\ \mathbf{0}^\mathsf{T} & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{I} & -\boldsymbol{X}_0 \\ \mathbf{0}^\mathsf{T} & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{X} \\ 1 \end{pmatrix} \quad . \tag{1}$$

With the projection matrix (K is the matrix $\boldsymbol{K}$ multiplied by an arbitrary scalar $\neq 0$) we finally end up with

$$\mathsf{P} = \mathsf{K}\boldsymbol{R}(\boldsymbol{I} \,|\, -\boldsymbol{X}_0) \qquad \text{and} \qquad \mathbf{x}' = \mathsf{P}\mathbf{X} \quad . \tag{2}$$

With P image coordinates can be predicted linearly from object coordinates. Due to its homogeneity (multiplication with an arbitrary scalar $\neq 0$ does not change the projection) the 3×4 matrix P has only 11 degrees of freedom (DOF).

2.3 Fundamental and Essential Matrix

The fundamental matrix describes the (projective) relative orientation of the image pair. We assume that we have n homologous points $\mathbf{x}'_i$ in the first image and $\mathbf{x}''_i$ in the second image. The projection matrices are $\mathsf{P}_1 = \mathsf{K}'\boldsymbol{R}'(\boldsymbol{I}\,|\mathbf{0})$ and $\mathsf{P}_2 = \mathsf{K}''\boldsymbol{R}''(\boldsymbol{I}\,| - \mathbf{T})$.

Assuming $\boldsymbol{R}' = \boldsymbol{I}$, this corresponds to the method of relative orientation of successive photographs To simplify the presentation, we transform the observed image points in the system of an ideal camera with projection matrix $(\boldsymbol{I}\,|\mathbf{0})$. We obtain the reduced image coordinates marked with the superscript index '^{k}' ${}^k\mathbf{x}' = \boldsymbol{R}'^{-1}\mathsf{K}'^{-1}\mathbf{x}'$ and ${}^k\mathbf{x}'' = \boldsymbol{R}''^{-1}\mathsf{K}''^{-1}\mathbf{x}''$. The condition, that the rays should intersect, implies the coplanarity of $\mathbf{x}'$, $\mathbf{x}''$, and $\mathbf{T}$. Presented in the same coordinate system we obtain

$$\begin{gathered} {}^k\mathbf{x}' \cdot ({}^k\mathbf{T} \times {}^k\mathbf{x}'') = {}^k\mathbf{x}'^\mathsf{T}\mathsf{S}_T\, {}^k\mathbf{x}'' = 0 \\ \text{with} \qquad \boldsymbol{S}_T = \boldsymbol{S}(\boldsymbol{T}) = \begin{pmatrix} 0 & -T_3 & T_2 \\ T_3 & 0 & -T_1 \\ -T_2 & T_1 & 0 \end{pmatrix} \quad . \end{gathered}$$

$\boldsymbol{S}(\boldsymbol{T})$ is a skew-symmetric matrix for the vector $\boldsymbol{T}$ with rank 2, which allows the vector product $\boldsymbol{V} = \boldsymbol{T} \times \boldsymbol{U}$ to be written as a matrix vector multiplication: $\boldsymbol{V} = \boldsymbol{S}_T\boldsymbol{U} = -\boldsymbol{S}_U\boldsymbol{T}$. Putting things together we end up with

$$\mathbf{x}'^\mathsf{T}(\mathsf{K}'^{-1})^\mathsf{T}\boldsymbol{R}'\mathsf{S}_T\,\boldsymbol{R}''^{-1}\mathsf{K}''^{-1}\mathbf{x}'' = 0 \quad .$$

This relation is linear in the image coordinates of both images, i.e., bilinear. With the 3×3 fundamental matrix the coplanarity equation as condition for the homology of image points can be represented in a simple and elegant way:

$$\mathsf{F} = (\mathsf{K}'^{-1})^\mathsf{T}\boldsymbol{R}'\mathsf{S}_T\,\boldsymbol{R}''^{-1}\mathsf{K}''^{-1} \quad \text{i.e.,} \quad \mathbf{x}'^\mathsf{T}\mathsf{F}\mathbf{x}'' = 0 \quad .$$

The latter equation has some important properties:

- Because it refers to the original measured data, there is no need for a reduction of the image coordinates. The reduction is contained in the fundamental matrix. The bilinear form is linear in the coefficients of the fundamental matrix. This allows for a direct determination from homologous points.
- For all points $\mathbf{x}''$ in the second image which lie on the straight line
$$\mathbf{l}'' = \mathsf{F}^\mathsf{T}\mathbf{x}' \quad ,$$
holds $\mathbf{l}''\mathbf{x}'' = 0$ and, therefore, the coplanarity condition. I.e., $\mathbf{l}''$ is the epipolar line of $\mathbf{x}'$ in the second image. It can be used to predict the geometrical location of $\mathbf{x}''$ in the second image in the form of a straight line. The computation can be based solely on F. There is no need to know the parameters of the orientation of the two cameras.

The 3×3 fundamental matrix has 9 elements. As S_T is of rank 2, the fundamental matrix is singular with rank 2. Because it is additionally homogenous, it has only 7 DOF. The condition $|\mathsf{F}| = 0$ has to be enforced, which is cubic in the parameters of F.

If calibration data is available, the fundamental matrix reduces to the essential matrix E and its bilinear form

$$\mathsf{E} = \mathsf{S}_T\,\boldsymbol{R}^{-1} \qquad\qquad {}^r\mathbf{x}'^\mathsf{T}\mathsf{E}\,{}^r\mathbf{x}'' = 0 \quad ,$$

with reduced image coordinates ${}^r\mathbf{x}' = \mathsf{K}'^{-1}\mathbf{x}'$ and ${}^r\mathbf{x}'' = \mathsf{K}''^{-1}\mathbf{x}''$. The essential matrix can be obtained from the fundamental matrix from $\mathsf{E} = \boldsymbol{K}''^\mathsf{T}\mathsf{F}\boldsymbol{K}'$.

3 TRIFOCAL TENSOR

The idea of the trifocal tensor and its linear computation was presented for the first time in (Hartley, 1994, Shashua, 1994). The computation of a consistent tensor was described in (Torr and Zisserman, 1997). (Faugeras and Papadopoulo, 1997) deals with constraints on the trifocal tensor. In photogrammetry (Förstner, 2000, Ressel, 2000, Theiss et al., 2000) have described the trifocal tensor and reported about experiments.

3.1 Trifocal Geometry from the Image Pair

When extending the image pair by another image we basically assume that all three projection centers are different. When they are additionally not collinear, they form the trifocal plane. This plane intersects the image planes in the three trifocal lines $\mathbf{t}_1$, $\mathbf{t}_2$, and $\mathbf{t}_3$, which comprise the epipoles $\mathbf{e}_{i,j}$ (cf. Figure 1). It is important to note that the three fundamental matrices F_{12}, F_{23}, and F_{31} are not independent. They have to comply with the following three conditions arising from the coplanarity of all projection centers and epipoles:

$$\mathbf{e}_{2,3}^\mathsf{T}\mathsf{F}_{12}\mathbf{e}_{1,3} = \mathbf{e}_{3,1}^\mathsf{T}\mathsf{F}_{23}\mathbf{e}_{2,1} = \mathbf{e}_{1,2}^\mathsf{T}\mathsf{F}_{31}\mathbf{e}_{3,2} = 0 \quad .$$

To see why this is true, observe that, e.g., the epipolar line of the epipole $\mathbf{e}_{1,3}$ in the second image is represented by $\mathsf{F}_{12}\mathbf{e}_{1,3}$. $\mathbf{e}_{1,3}$ is the image point of O''' in the first image just as $\mathbf{e}_{2,3}$ in the second image. From the latter follows the first condition.

The orientation based on image triplets has some general advantages over the orientation based on image pairs:

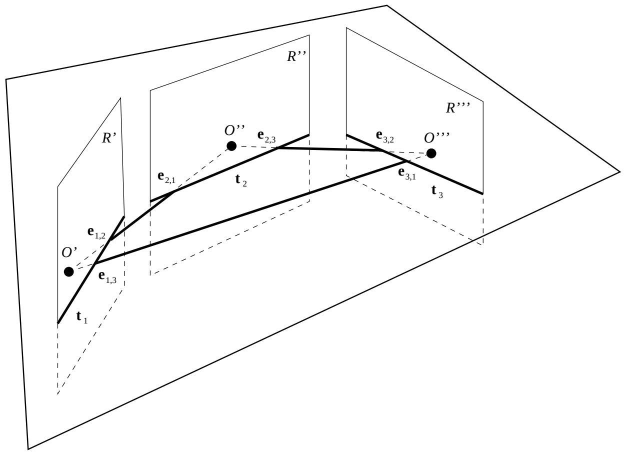

Figure 1: The trifocal plane

- The orientation can be based upon homologous points in the same way as on (infinitely long) homologous straight lines.
- Practical experience shows that the local geometry of an image strip or an image sequence can be much more precisely and, what is more important, also much more robustly determined from image triplets and their conditions than from only weakly overdetermined pairs. This is true for the trifocal tensor but also for bundle triangulation. Opposed to the latter, the trifocal tensor has like the fundamental matrix its strength in its linearity. Linearity equals speed and this makes the determination of approximate solutions, e.g., based on RANSAC (cf. Section 5.2) possible.

3.2 Derivation of the Trifocal Tensor

The trifocal tensor can be introduced intuitively based on homologous straight lines (Hartley and Zisserman, 2000). Given are a straight line $\mathbf{l}'$ in the first and a straight line $\mathbf{l}''$ in the second image (cf. Figure 2). The planes $\boldsymbol{\pi}' = \mathsf{P}'^{\mathsf{T}}\mathbf{l}''$ and $\boldsymbol{\pi}'' = \mathsf{P}''^{\mathsf{T}}\mathbf{l}''$ constructed from these lines intersect in the 3D straight line L, the image of which in the third camera is in general the straight line $\mathbf{l}'''$.

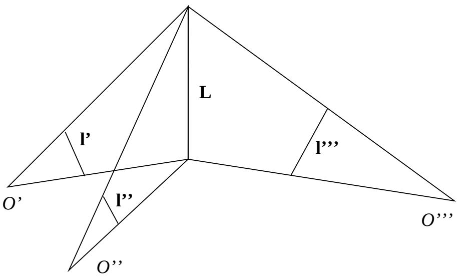

Figure 2: Trifocal tensor from three intersecting lines

For the fundamental matrix as well as for the trifocal tensor the projection matrices can always be transformed in a way that the projection matrix of the first image is $\mathsf{P}' = [\mathsf{I}|\mathbf{0}]$. The other two projection matrices can then be written as $\mathsf{P}'' = [\mathsf{A}|\mathbf{a}_4]$ and $\mathsf{P}''' = [\mathsf{B}|\mathbf{b}_4]$. Based on this the intersection of the three lines in 3D space can be defined algebraically by requiring that the 4×3 matrix $\mathsf{M} = [\boldsymbol{\pi}', \boldsymbol{\pi}'', \boldsymbol{\pi}''']$ has rank 2. Points on the line of intersection may be represented as $\mathbf{X} = \alpha\,\mathbf{X}_1 + \beta\,\mathbf{X}_2$ with $\mathbf{X}_1$ and $\mathbf{X}_2$ linearly independent. These points are incident to all three planes and thus $\boldsymbol{\pi}'^{\mathsf{T}}\mathbf{X} = \boldsymbol{\pi}''^{\mathsf{T}}\mathbf{X} = \boldsymbol{\pi}'''^{\mathsf{T}}\mathbf{X} = 0$. This implies $\mathsf{M}^{\mathsf{T}}\mathbf{X} = 0$ and because of $\mathsf{M}^{\mathsf{T}}\mathbf{X}_1 = 0$ and $\mathsf{M}^{\mathsf{T}}\mathbf{X}_2 = 0$ M has a 2D null-space.

Since the rank of M is 2, there is a linear dependence among the columns. If we denote

$$\mathsf{M} = [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3] = \begin{bmatrix} \mathbf{l}' & \mathsf{A}^{\mathsf{T}}\mathbf{l}'' & \mathsf{B}^{\mathsf{T}}\mathbf{l}''' \\ 0 & \mathbf{a}_4^{\mathsf{T}}\mathbf{l}'' & \mathbf{b}_4^{\mathsf{T}}\mathbf{l}''' \end{bmatrix} \quad ,$$

the linear relation may be written $\mathbf{m}_1 = \alpha\mathbf{m}_2 + \beta\mathbf{m}_3$. Because the lower left element of M is 0 it follows that $\alpha = k(\mathbf{b}_4^{\mathsf{T}}\mathbf{l}''')$ and $\beta = -k(\mathbf{a}_4^{\mathsf{T}}\mathbf{l}'')$ for some scalar k. Applying this to the upper three vectors we obtain up to a homogeneous scale factor

$$\mathbf{l} = (\mathbf{b}_4^{\mathsf{T}}\mathbf{l}''')\mathsf{A}^{\mathsf{T}}\mathbf{l}'' - (\mathbf{a}_4^{\mathsf{T}}\mathbf{l}'')\mathsf{B}^{\mathsf{T}}\mathbf{l}''' = (\mathbf{l}'''^{\mathsf{T}}\mathbf{b}_4)\mathsf{A}^{\mathsf{T}}\mathbf{l}'' - (\mathbf{l}''^{\mathsf{T}}\mathbf{a}_4)\mathsf{B}^{\mathsf{T}}\mathbf{l}''' \ .$$

The i-th coordinate of $\mathbf{l}'$ can thus be written

$$\mathbf{l}'_i = \mathbf{l}'''^{\mathsf{T}}(\mathbf{b}_4\mathbf{a}_i^{\mathsf{T}})\mathbf{l}'' - \mathbf{l}''^{\mathsf{T}}(\mathbf{a}_4\mathbf{b}_i^{\mathsf{T}})\mathbf{l}''' = \mathbf{l}''^{\mathsf{T}}(\mathbf{a}_i\mathbf{b}_4^{\mathsf{T}})\mathbf{l}''' - \mathbf{l}''^{\mathsf{T}}(\mathbf{a}_4\mathbf{b}_i^{\mathsf{T}})\mathbf{l}''' \ .$$

By denoting $\mathsf{T}_i = \mathbf{a}_i\mathbf{b}_4^{\mathsf{T}} - \mathbf{a}_4\mathbf{b}_i^{\mathsf{T}}$ the incidence relation can finally be written as

$$\mathbf{l}'_i = \mathbf{l}''^{\mathsf{T}}\mathsf{T}_i\mathbf{l}''' \quad . \qquad (3)$$

T is a bilinear transformation and defines the trifocal tensor which is usually written as $\mathcal{T}_i^{jk}$. It is a $3 \times 3 \times 3$ cube made up of 27 elements.

- $\mathcal{T}_i^{jk}$ has 18 DOF at maximum. I.e., not all cubes are trifocal tensors. The number 18 is obtained by subtracting from the 33 parameters of the three projection matrices the 15 parameters for a projective transformation (homogenous 4×4 matrix) of space. For the solution either the conditions have to be enforced or the trifocal tensor has to be minimally parameterized. Both lead to non-linear equations. Practical investigations have shown, that it is important to use the conditions, because the solution is not stable otherwise.
- There are direct relations among the coefficients of the trifocal tensor, the fundamental matrices of the three image pairs, and the three projection matrices. They can be employed to determine from the trifocal tensor the fundamental matrices and after the choice of a coordinate system also the projection matrices.

4 POINT TRANSFER WITH THE TRIFOCAL TENSOR

Based on the trifocal tensor a prediction of points and straight lines in the third image is feasible without determining the point or the straight line in space. I.e., the trifocal tensor describes relations between measurements in the images *without the need to reconstruct 3D geometry explicitly*. In principle this corresponds to the epipolar line for the image pair, but opposed to it the result is unique.

For the general case the prediction for points could be done by intersecting the epipolar lines in the third image corresponding to the homologous points in the first and the second image, respectively. This is true only if the epipolar lines are not parallel, which is the case if a point lies on the trifocal plane, or if the projection centers are collinear. The latter is often valid or at least nearly valid, e.g., for aerial images from one flight strip.

The restriction of the preceding paragraph does not hold if we employ the trifocal tensor: Given two homologous points $\mathbf{x}'$ and $\mathbf{x}''$ in the first and the second image one chooses a line $\mathbf{l}''$ through $\mathbf{x}''$. Then, the point $\mathbf{x}'''$ can be computed by transferring $\mathbf{x}'$ from the first to the third view via the homography defined by $l''_j \mathcal{T}_i^{jk}$, i.e., $x'''^k = x'^i l''_j \mathcal{T}_i^{jk}$.

This transfer via the homography implies the intersection of the plane defined by the projection center of the second camera O'' and the line $\mathbf{l}''$ with the ray defined by the projection center of the first camera O' and $\mathbf{x}'$ (cf. Fig. 3). This intersection is not defined if $\mathbf{l}''$ is taken to be the epipolar line corresponding to $\mathbf{x}'$. The plane defined by the epipolar line and O'' comprises the point $\mathbf{X}$ which is projected to $\mathbf{x}'$, $\mathbf{x}''$, and $\mathbf{x}'''$ as well as the projection center O' of the first camera. In this plane lies also the ray defined by O' and $\mathbf{x}'$.

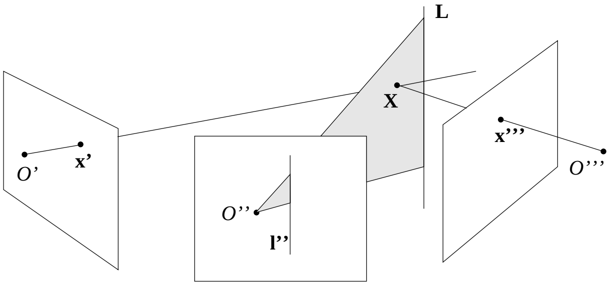

Figure 3: Transfer of a point $\mathbf{x}'$ in the first image via a plane defined by the line $\mathbf{l}''$ in the second image and the projection center of the second camera O''.

On the other hand, if one takes the line perpendicular to the epipolar line through $\mathbf{x}''$, also the projection plane becomes in one direction perpendicular to the ray defined by $\mathbf{x}'$ and O' and thus the intersection geometry becomes optimal. In (Hartley and Zisserman, 2000) it is recommended to use optimal triangulation to compute a pair $\mathbf{x}'$ and $\mathbf{x}''$ which satisfies $\mathbf{x}''^{\mathsf{T}} \mathsf{F}_{12} \mathbf{x}' = 0$. We do not do this because we are focusing on speed rather than on optimum quality. Putting everything together, our basic algorithm looks as follows:

- Compute $\mathbf{l}''$ which goes through $\mathbf{x}''$ and is perpendicular to $\mathbf{l}''_e = \mathsf{F}_{21}\mathbf{x}'$. From $\mathbf{x}'' = (x''_1, x''_2, 1)^{\mathsf{T}}$ and $\mathbf{l}''_e = (l''_1, l''_2, l''_3)^{\mathsf{T}}$ follows $\mathbf{l}'' = (l''_2, -l''_1, -x''_1 l''_2 + x''_2 l''_1)^{\mathsf{T}}$
- The transfered point is $x'''^k = x'^i l''_j \mathcal{T}_i^{jk}$.

A closer look at this reveals that if one restricts oneself to points on the epipolar line, then only $l''_3 = -x''_1 l''_2 + x''_2 l''_1$ varies. Therefore, $x''^i l''_1 \mathcal{T}_i^{1k}$ and $x''^i l''_2 \mathcal{T}_i^{2k}$ are constant and have to be computed only once per epipolar line.

5 ESTIMATION OF THE TRIFOCAL TENSOR

5.1 Carlsson-Weinshall Duality

We employ the Carlsson-Weinshall duality to calculate a solution for the trifocal tensor from a minimum of six point triplets (Carlsson, 1995, Weinshall et al., 1995). To utilize an algorithm which gives a solution for a minimum number of points is important in two ways: First, for robust estimation based, e.g., on RANSAC (cf. Section 5.2), this considerably reduces the search space. Second, by taking the minimum number of points we implicitly take the constraints for a tensor to be a trifocal tensor into account.

The basic idea of the duality is to interchange the roles of points being viewed by several cameras and the projection centers. Specifically, if one has an algorithm for n views and $m+4$ points, then there is an algorithm for doing projective reconstruction from m views of $n+4$ points. By taking into account $|\mathsf{F}| = 0$ (cf. Section 2.3), an algorithm can be constructed for the reconstruction of the fundamental matrix from two images for which seven homologous points suffice. From the above follows that $m = 3$ and $n = 2$. I.e., if we utilize the dualism we get an algorithm solving for three images and six points.

To determine the fundamental matrix from seven points, we start with the basic solution for $n \geq 8$ homologous points. For it the homogenous linear equation system reads $\mathbf{x}'^{\mathsf{T}}_i \mathsf{F} \mathbf{x}''_i = 0$, i.e., $\mathsf{A}_i \mathbf{u} = \mathbf{0} \quad \forall i = 1, \ldots, n$. With the 9-vector $\mathbf{u}$ representing the elements of F this can be written in the form $\mathsf{Au} = \mathbf{0}$ with $\mathsf{A} = (\mathsf{A}_i)$. The best estimation for $\mathbf{u}$ is the unit singular vector corresponding to the smallest singular value of the $n \times 9$-matrix A, determined by singular value decomposition (SVD). The solution is unique up to an unknown factor, as the system is homogenous. For seven points the 7×9-matrix A has rank 7. The solution for $\mathsf{Au} = \mathbf{0}$ is a 2D space with the form $\alpha \mathsf{F}_1 + (1-\alpha)\mathsf{F}_2$. Using the fact that F is of rank 2 leads to $|\alpha \mathsf{F}_1 + (1-\alpha)\mathsf{F}_2)| = 0$. This results into a cubic polynomial for which either one or three real solutions exist.

The dual algorithm is described in detail in (Hartley and Zisserman, 2000). Here we only sketch it. The triplets of points of the original problem are arranged in a table and the table is transformed so that the last four points are mapped to the 2D projective basis, i.e., $(1,0,0)^{\mathsf{T}}$, $(0,1,0)^{\mathsf{T}}$, etc. Then, the last four points are dropped, the table is transposed and extended by points of the 2D projective basis. The solution for the dual problem is obtained by the algorithm for the fundamental matrix. The obtained reconstruction is transformed in a way that the last four points correspond to the 3D projective basis. By dualization the problem is mapped back into the original domain. Finally, the effects of the initial transformation are undone by a reverse transformation.

5.2 Dealing with Mismatches by RANSAC

Even though there are generic means to reduce mismatches (cf. Section 5.3), there are usually far too many for an efficient least squares solution, if the knowledge about the calibration and the orientation of the cameras is weak. As our problem is of the type that we only have relatively few parameters and a high redundancy, the RANSAC (random sample consensus) approach (Fischler and Bolles, 1981) is a good choice. RANSAC is based on the idea to select more or less randomly minimum sets of observations. The correctness of a set is evaluated by the number of other observations which confirm it.

The minimum number of point correspondences necessary to determine the trifocal tensor with RANSAC is seven for the fundamental matrix and six for the trifocal tensor (cf. Section 5.1). At first sight the number for the trifocal tensor looks better than that for the fundamental matrix. Yet, one has to consider, that there exist 6^3 combinations for six triplets, which is considerably more than the 7^2 for seven pairs. This suggests, that we first calculate the correspondences based on the fundamental matrices of the images one and two as well as one and three and only then match the triplets.

5.3 Reduction of the Search Space and Results

By means of a hierarchical approach based on image pyramids with a reduction by a factor 2 for each level, we significantly reduce the search space. Thereby not only the efficiency but also the robustness is improved considerably.

Highly precise conjugate points are obtained from a least-squares matching of points obtained from the sub-pixel Förstner operator (Förstner and Gülch, 1987). On the highest level of the pyramids which usually consist of about 100×100 pixels, no reduction of the search space, e.g., by means of epipolar lines, is yet available. To reduce the complexity of the matching, several measures are taken. First, before the actual least-square matching we sort out many points and calculate a first approximation by thresholding and maximizing, respectively, the correlation score among image windows. What is more, we restrict ourselves in the first image to only a few hundred of the globally strongest points and some more points which are strongest regionally on a even-spaced grid.

Section 4 made clear that the trifocal tensor is superior for point transfer compared to the intersection of epipolar lines. Yet, before one has an approximation of the trifocal tensor it is a good idea to compute fundamental matrices to narrow down the search space (cf. Section 5.2). We do this in two ways: On the highest pyramid level we compute the fundamental matrices and from them the epipolar lines from the first to the second and to the third image before we actually search for image triplets. After we have obtained an approximation of the trifocal tensor, we compute from it the fundamental matrix from the first to the second image on the lower levels. For the points found by matching on the epipolar line in the second image we predict their position in the third image. Figure 4 shows the first image with an extracted point on the left side. On the epipolar line in the second image two points were found. Those two lead in turn to the prediction of the two points in the third image for which one is obviously wrong and would not be matched.

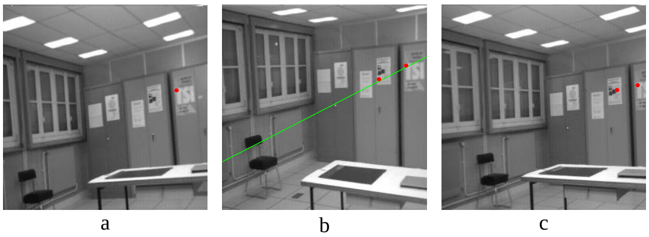

Figure 4: Prediction: From the point in a) the epipolar line in b) arises from the fundamental matrix. The two points found in b) uniquely determine the two points in c) via the trifocal tensor.

The linear solution for the trifocal tensor is based on RANSAC (cf. Section 5.2). To improve the results, the parameters of the second and third cameras are finally optimized by the non-linear Levenberg-Marquardt algorithm (Hartley and Zisserman, 2000). The approach was implemented in C^{++} making use of the public domain linear algebra package LAPACK interfaced by the template numerical toolkit (TNT; math.nist.gov/tnt). Results are shown in Figures 5 and 6. Please note that we use by default a quadratic search space which covered in the case of both Figures 50% of the image area.

6 DEPTH ESTIMATION

Our depth estimation, which can deal with strong occlusions and large disparity ranges also for details, i.e., very tiny structures in the image, is built upon the approach proposed in (Zitnick and Kanade, 2000). It employs epipolar resampled imagery. The basic idea is to calculate matching scores for a disparity range and store this information in a 3D array made up of image width and height as well as disparity range. This array is then filtered with a 3D box-filter to obtain the local support for a match from all close-by matches. On the other hand, it is assumed that on one ray of view only one point is visible. This implies an inhibition which is realized by weighting down all scores besides the strongest. Support and inhibition are iterated. Thereby, the information is propagated more globally. To avoid hallucination, the original matching scores are considered after each iteration. Finally, occlusion regions are marked by thresholding the matching scores. We have extended this approach with the following features:

- Additionally to normalized cross-correlation we employ absolute differences as proposed by (Scharstein and Szeliski, 2002) for the matching scores.
- We automatically estimate the disparity range from a number of sample lines. This works relatively robustly and considerably reduces the search space and therefore the computation time.
- By separating the 3D box-filter into 2D planes and 1D sticks we have sped up the computation for this part by a factor of 5.
- We determine the convergence of the algorithm automatically by calculating a difference image and setting a threshold on its mean and variance.
- The smoothness of the output is improved by a sub-pixel disparity estimation in the original matching scores.

7 VIEW SYNTHESIS WITH THE TRIFOCAL TENSOR

We use the view synthesis scheme proposed in (Avidan and Shashua, 1998). The basic idea is to use calibrated imagery together with a depth map. With the latter, points homologous to points in the first image are obtained for the second image.

At least a weak calibration is necessary to make a navigation through the image meaningful for the user as only then rotation matrices and translation vectors are defined in a Euclidean sense. If calibration information is available, we compute $\mathbf{E}$ from it. It can be separated into a rotation matrix $\boldsymbol{R}$ and a translation vector $\boldsymbol{t}$ via SVD. Together they make up the projection matrix $\mathbf{P}'' = [\boldsymbol{R}|\boldsymbol{t}]$ if we assume $\mathbf{P}' = [\boldsymbol{I}|\mathbf{0}]$. To improve the results, the three parameters of the rotation matrix and the two parameters of the translation vector are optimized by Levenberg-Marquardt. If no calibration information is available, we assume that both images are taken with the same camera without modifications. We further assume that the principal point equals the image center, that there is no sheer, i.e., $s = 0$, that the focal length is one and that the pixels are square. We provisionally calibrate the fundamental matrix with these assumptions and when optimizing by Levenberg-Marquardt we vary two additional parameters describing c as well as the relation of the width and the height of a pixel.

The trifocal tensor is initially instantiated from the fundamental matrix

$$\mathcal{T}_i^{jk} = \epsilon^{ljk} F_{li} \quad ,$$

where ϵ^{ljk} is the cross-product tensor and F_{li} is $\mathbf{F}$ in tensor notation. Then, the view synthesis is accomplished by modifying the trifocal tensor by rotation matrices $\boldsymbol{R}$ (R_i^j in tensor notation) and translation vectors t given by the user. The modified tensor is

$$\mathcal{G}_i^{jk} = R_l^k \mathcal{T}_i^{jl} + t^k a_i^j \quad ,$$

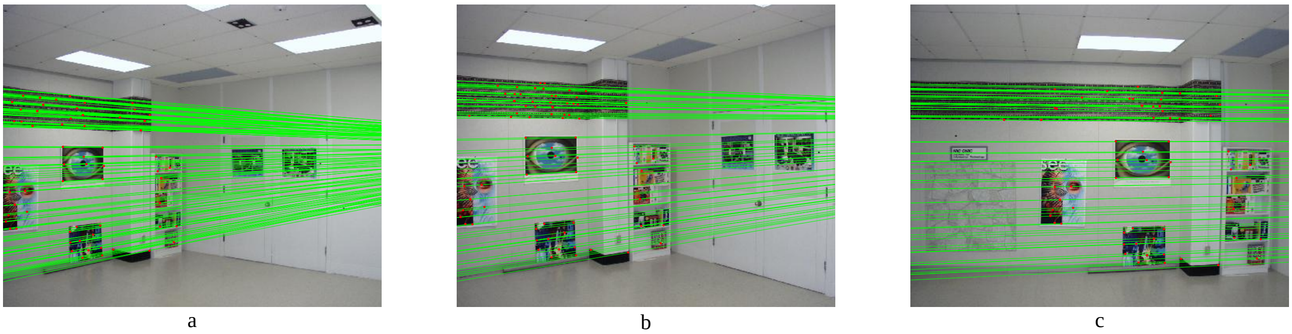

Figure 5: Result I: Image triplet with points and epipolar lines. a) Epipolar lines for b) and c) b) and c) Epipolar lines for a) only

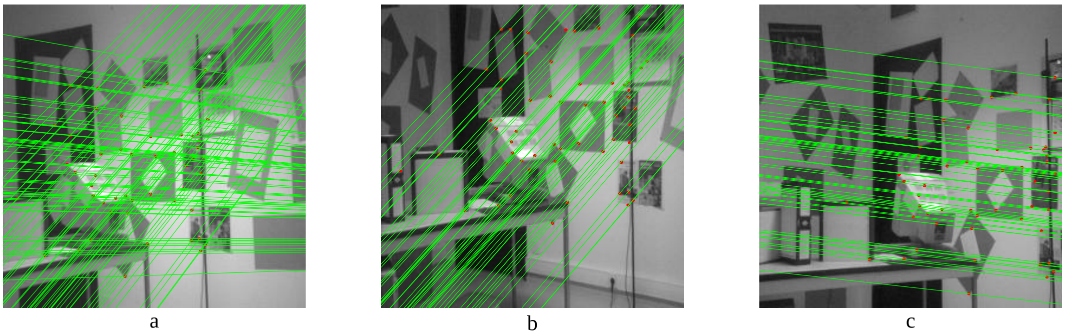

Figure 6: Result II: Image triplet with points and epipolar lines. a) Epipolar lines for b) and c) b) and c) Epipolar lines for a) only

where a_i^j is the first part $\mathbf{A}$ of the calibrated projection matrix of the second camera.

The actual projection is done based on the optimized scheme proposed in Section 4. The synthesized image is produced indirectly by mapping the pixels via affine transformation obtained by the known coordinates of triangle meshes in the given and the synthesized image. Results are shown in Figures 7, 8, and 9. In all cases only the translation vectors have been modified. The synthesized views give an impression of the depth in the image. Compared to approaches for view interpolation also the quality for viewing directions which are not in the direction of the base vector $\mathbf{T}$ is reasonable (cf., e.g., Figure 9 for which the epipolar lines are nearly vertical). On the other hand, one can see the problems arising from the height data. Besides the fact that parts had too bad texture to be matched (black holes in the depth map), also the boundary of structures with large disparities such as the chair in Figure 8 are not well delineated.

8 CONCLUSIONS

In this paper we have presented means for the estimation of the trifocal tensor as well as its use for view synthesis. All results presented have been obtained totally automatically, without any user interaction. The same parameters have been used for all examples. While the estimation of the trifocal tensor based on pyramids, least squares matching, and RANSAC works reliable for a wide range of imagery, the end-to-end automation of view synthesis still is an intricate problem. There are two things we still have to cope with in-depth: Camera calibration and depth estimation.

For camera calibration we have begun to implement the approach presented in (Pollefeys and Van Gool, 1999, Hartley and Zisserman, 2000). The more complicated problem is depth estimation. Opposed to the determination of the orientation which is defined by very few parameters and is therefore a highly redundant problem, depth estimation aims at determining many parameters. Although the approach we are using is relatively sophisticated, the results are in many instances unstable and not really good. One way for improvement would be to use more images. This makes it computationally much more expensive as the simple epipolar geometry cannot be used any more. Other ways for improvement were recently shown in (Scharstein and Szeliski, 2002). As the most important problem is the determination of approximate values, a combination with direct sensors with possibly a lower resolution such as cheap laser-scanners planned, e.g., for airbag inflation control, might be considered for the application domain of video communication.

ACKNOWLEDGMENTS

The images used in Figures 4, 6, 8, and 9 are courtesy of the Robotvis group at INRIA and the images utilized in Figures 5 and 7 stem from ISPRS Working Group V/2.

We thank Peter Krzystek for making us available his code for least squares matching.

REFERENCES

Avidan, S. and Shashua, A., 1998. Novel View Synthesis by Cascading Trilinear Tensors. IEEE Transactions on Visualization and Computer Graphics 4(4), pp. 293–306.

Brandstätter, G., 1996. Fundamentals of Algebro-Projective Photogrammetry. In: Sitzungsberichte, Mathematisch-naturwissenschaftliche Klasse Abt. II, Mathematische, Physikalische und Technische Wissenschaften, Österreichische Akdademie der Wissenschaften, Vol. II (1996) 205, pp. 57–109.

Carlsson, S., 1995. Duality of Reconstruction and Positioning from Projective Views. In: IEEE Workshop on Representation of Visual Scenes, Boston, USA.

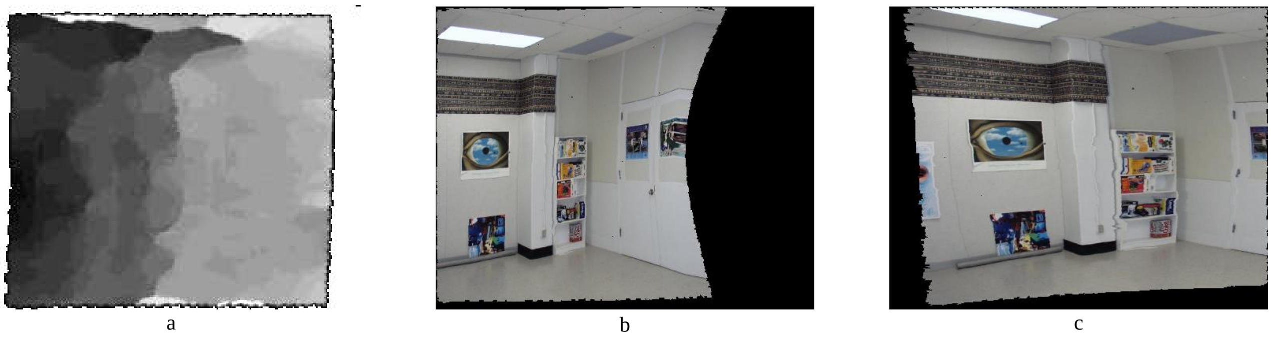

Figure 7: Results I: a) Depth map (white: not enough texture). b) and c) synthesized views

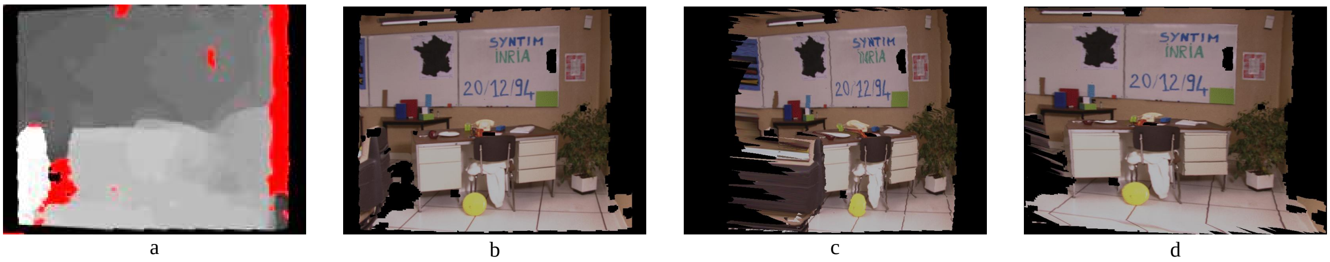

Figure 8: Results II: a) Depth map image (white: not enough texture; red: occlusion regions). b), c), and d) synthesized views

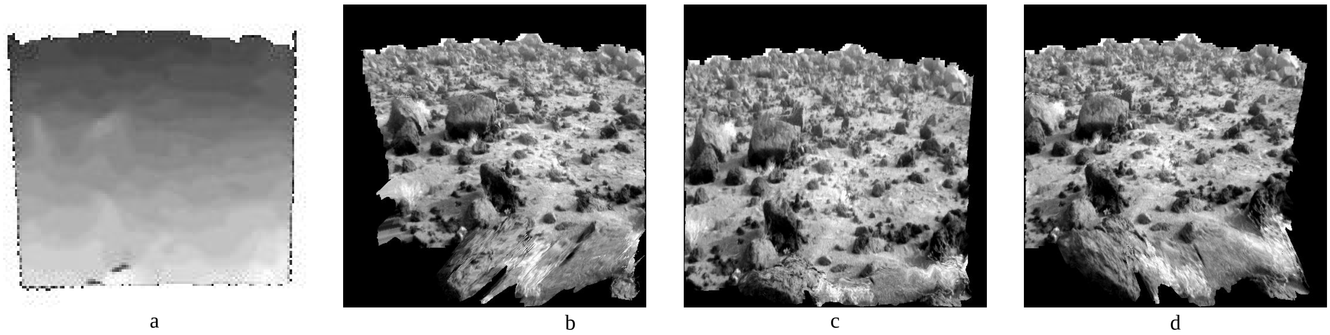

Figure 9: Results III: a) Depth map (white: not enough texture). b), c), and d) synthesized views

Faugeras, O. and Papadopoulo, T., 1997. Grassman-Cayley Algebra for Modeling Systems of Cameras and the Algebraic Equations of the Manifold of Trifocal Tensors. Rapport de Recherche 3225, INRIA, Sophia Antipolis, France.

Fischler, M. and Bolles, R., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Communications of the ACM 24(6), pp. 381–395.

Förstner, W., 2000. New Orientation Procedures. In: International Archives of Photogrammetry and Remote Sensing, Vol. (33)B3/1, pp. 297–304.

Förstner, W. and Gülch, E., 1987. A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features. In: ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data, Interlaken, Switzerland, pp. 281–305.

Hartley, R., 1994. Projective Reconstruction from Line Correspondence. In: Computer Vision and Pattern Recognition, pp. 903–907.

Hartley, R. and Zisserman, A., 2000. Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge, UK.

Pollefeys, M. and Van Gool, L., 1999. Stratified Self-Calibration with the Modulus Constraint. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(8), pp. 707–724.

Ressel, C., 2000. An Introduction to the Relative Orientation Using the Trifocal Tensor. In: International Archives of Photogrammetry and Remote Sensing, Vol. (33) B3/2, pp. 769–776.

Scharstein, D. and Szeliski, R., 2002. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. International Journal of Computer Vision 47(1), pp. 7–42.

Shashua, A., 1994. Trilinearity in Visual Recognition by Alignment. In: Third European Conference on Computer Vision, Vol. I, pp. 479–484.

Theiss, H., Mikhail, E., Aly, I., Bethel, J. and Lee, C., 2000. Photogrammetric Invariance. In: International Archives of Photogrammetry and Remote Sensing, Vol. (33) B3/2, pp. 584–591.

Torr, P. and Zisserman, A., 1997. Robust Parametrization and Computation of the Trifocal Tensor. Image and Vision Computing 15, pp. 591–605.

Weinshall, D., Werman, M. and Shashua, A., 1995. Shape Descriptors: Bilinear, Trilinear, and Quadrilinear Relations for Multi-Point Geometry and Linear Projective Reconstruction. In: IEEE Workshop on Representation of Visual Scenes, Boston, USA, pp. 55–65.

Zitnick, C. and Kanade, T., 2000. A Cooperative Algorithm for Stereo Matching and Occlusion Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(7), pp. 675–684.