FAST TRIFOCAL TENSOR ESTIMATION USING VIRTUAL PARALLAX †

Manolis I. A. Lourakis and Antonis A. Argyros

Institute of Computer Science, Foundation for Research and Technology - Hellas (FORTH)
Vassilika Vouton, P.O. Box 1385, GR 711 10, Heraklion, Crete, GREECE
{lourakis, argyros}@ics.forth.gr http://www.ics.forth.gr/cvrl/

ABSTRACT

We present a computationally efficient method for estimating the trifocal tensor corresponding to three images acquired by a freely moving camera. The proposed method represents projective space through a "plane + parallax" decomposition and employs a novel technique for estimating the homographies induced by a virtual 3D plane between successive image pairs. Knowledge of these homographies allows the corresponding camera projection matrices to be expressed in a common projective frame and, therefore, to be recovered directly. The trifocal tensor can then be recovered in a straightforward manner from the estimated projection matrices. Sample experimental results demonstrate that the method performs considerably faster compared to a state of the art method, without a serious loss in accuracy.

1. INTRODUCTION

The trifocal tensor encapsulates all geometric relations among three images and arises as a relationship between corresponding point and line features, independently of the scene structure. For this reason, the trifocal tensor has been employed in a wide variety of vision tasks, including, among others, projective and euclidean 3D reconstruction, camera tracking, recognition by alignment, novel view synthesis and object-based video compression. Owing to its wide applicability, the problem of estimating the tensor from corresponding image features in three views has received considerable attention. Among the numerous relevant publications, the works of [1, 2] are the most representative ones. In [1], Hartley proposes a linear algorithm that combines corresponding points and lines in three views by exploiting the fact that each point and line correspondence provides respectively four and two linear equations in the elements of the tensor. Torr and Zisserman [2] describe an iterative non-linear algorithm for computing a Maximum Likelihood Estimate (MLE) of the tensor. Chapter 15 of [3] describes an algorithm which incorporates more recent developments to the approach of [2] and represents the state of the art in feature-based tensor estimation. As a general comment on existing methods, linear ones perform faster and are simpler to implement. However, owing to the fact that they minimize algebraic error terms with no physical meaning, they are less accurate compared to non-linear methods. A related shortcoming of linear methods is that they do not enforce the nonlinear constraints that must be fulfilled by a valid tensor. For these reasons, linear methods are typically employed for bootstrapping more elaborate but computationally expensive non-linear ones.

In this work, a novel and fast feature-based approach to trifocal tensor estimation is presented. Image motion is expressed as the sum of a homographic transfer plus a residual planar parallax vector and a "chaining" operation is employed, which propagates across successive frames the image-to-image homographies induced by a virtual 3D plane. It is well-known that knowledge of a planar homography and of the associated epipole permits the camera projection matrices of the underlying image pair to be recovered directly up to an arbitrary projective transformation. In our case, the homographies pertaining to the virtual plane are estimated in such a way that all projection matrices recovered from them are, by construction, defined up to the same projective transformation. Such a set of consistent projection matrices yields the trifocal tensor in a straightforward, direct manner. The rest of the paper is organized as follows. Section 2 explains the notation that will be used and provides some background knowledge along with a means for selecting a virtual 3D plane. Section 3 describes plane homography chaining and section 4 builds upon it for dealing with the tensor estimation problem. Sample experimental results are reported in section 5. The paper concludes with a brief discussion in section 6.

2. NOTATION AND BACKGROUND

In the following, vectors and arrays appear in boldface and are represented using projective (homogeneous) coordinates; $\simeq$ denotes equality up to an arbitrary scale factor. 3D points are written in uppercase and their image projections in lowercase (e.g. $\mathbf{X}$ and $\mathbf{x}$). $\mathbf{F}$ will designate the fundamental matrix while $\mathbf{e}$ and $\mathbf{e}'$ its associated epipoles.

As shown in [4], the fundamental matrix and plane homographies are tightly coupled. Specifically, the group of

† PARTIALLY SUPPORTED BY THE EU FP6-507752 NOE MUSCLE.

0-7803-9134-9/05/$20.00 ©2005 IEEE

all possible homography matrices between two images lies in a subspace of dimension 4, i.e. it is spanned by 4 homography matrices. These homography matrices are such that their respective planes do not all coincide with a single point. Shashua and Avidan show in [5] that given the fundamental matrix $\mathbf{F}$ and the epipoles $\mathbf{e}$ and $\mathbf{e}'$ in an image pair, a suitable basis of 4 homography matrices $\mathbf{H}_1, \ldots, \mathbf{H}_4$, referred to as "primitive homographies", is defined as follows:

$$\mathbf{H}_i = [\epsilon_i]_\times \mathbf{F}, \quad i = 1, 2, 3 \qquad \text{and} \qquad \mathbf{H}_4 = \mathbf{e}' \delta^T, \quad (1)$$

where ϵ_i are the identity vectors $\epsilon_1 = (1, 0, 0), \epsilon_2 = (0, 1, 0)$ and $\epsilon_3 = (0, 0, 1)$, $[.]_\times$ designates the skew symmetric matrix representing the vector cross product and δ is a vector such that $\delta^T \mathbf{e} \neq 0$. Knowledge of the 4 primitive homographies allows any other homography $\mathbf{H}$ to be expressed as a linear combination $\mathbf{H} = \sum_{i=1}^{4} \lambda_i \mathbf{H}_i, \lambda_i \in \mathcal{R}$.

Next, a result due to Shashua and Navab [6] that plays a central role in the development of the proposed method is presented. Let Π be an arbitrary 3D plane inducing a homography $\mathbf{H}$ between two images. Let also $\mathbf{X}_0$ be a 3D point not on Π projecting to image points $\mathbf{x}_0$ and $\mathbf{x}_0'$ and assume that $\mathbf{H}$ has been scaled to satisfy the equation $\mathbf{x}_0' \simeq \mathbf{H}\mathbf{x}_0 + \mathbf{e}'$. Then, for any 3D point $\mathbf{X}$ projecting onto $\mathbf{x}$ and $\mathbf{x}'$, there exists a scalar κ such that

$$\mathbf{x}' \simeq \mathbf{H}\mathbf{x} + \kappa \mathbf{e}'. \quad (2)$$

Equation (2) dictates that the position of projected points in the second image can be decomposed into the sum of two terms, the first depending on the homography induced by Π and the second involving *parallax* due to the deviation of the actual 3D structure from Π. The term κ in Eq. (2) depends on $\mathbf{X}$ but is invariant to the choice of the second image and is termed as *relative affine structure* in [6]. Recall that $\mathbf{H}$ and $\mathbf{e}'$ are homogeneous entities, defined up to an arbitrary scale factor. Therefore, by fixing $\mathbf{H}$'s scale, the role of point $\mathbf{X}_0$ in the derivation of Eq. (2) is to establish a common relative scale between $\mathbf{H}$ and $\mathbf{e}'$. Notice, however, that in the case that $\mathbf{H}$ has not been scaled with the aid of $\mathbf{X}_0$, Eq. (2) continues to hold for some κ' that is a scaled version of that corresponding to the scaled $\mathbf{H}$. Given $\mathbf{x}$, $\mathbf{x}'$, $\mathbf{H}$ and $\mathbf{e}'$, the term κ corresponding to $\mathbf{X}$ can be computed by cross-multiplying both sides of Eq. (2) with $\mathbf{x}'$, which after some algebraic manipulation yields

$$\kappa = \frac{(\mathbf{H}\mathbf{x} \times \mathbf{x}')^T (\mathbf{x}' \times \mathbf{e}')}{||\mathbf{x}' \times \mathbf{e}'||^2}. \quad (3)$$

A final point in this section that will be needed in subsequent developments concerns the selection of a virtual 3D plane and its homography from a set of matching point pairs $(\mathbf{x}_i, \mathbf{x}_i')$, $i = 1, \ldots, N$ in two frames. The virtual plane can be chosen so that it approximates the set of available point matches as much as possible. In other words, the virtual plane is selected to lie "in-between" the 3D points giving rise to the set of available point matches. Assuming that the epipolar geometry corresponding to the two images has been estimated, we therefore seek the planar homography $\mathbf{U}$ for which the contribution of the parallax term in Eq. (2) is minimal. As previously explained, any planar homography defined between two images can be expressed as the linear combination of the four primitive homographies of Eq. (1). The sought $\mathbf{U}$ is thus computed from the coefficients μ_j, $j = 1, \ldots, 4$ minimizing $\sum_{j=1}^{4} (\mu_j \mathbf{H}_j) \mathbf{x}_i \simeq \mathbf{x}_i'$, $i = 1, \ldots, N$. This minimization is performed using the Least Median of Squares (LMedS) robust estimator and gives rise to a plane that is referred to as "quasi-metric" in [5].

3. RELATING HOMOGRAPHIES AMONG IMAGES

Assume that N triplets of matching points $(\mathbf{x}_i, \mathbf{x}_i', \mathbf{x}_i'')$, $i = 1, \ldots, N$, are available across three consecutive images I_1, I_2 and I_3 respectively and that the homography $\mathbf{U}$ of the quasi-metric plane Π_q for images I_1 and I_2 has been estimated as explained in the previous section [1]. This section describes a procedure for estimating the plane homography $\mathbf{V}$ induced by Π_q between images I_2 and I_3, without segmenting the point triplets into planar and non-planar ones or projectively reconstructing 3D points.

The epipolar geometry for images I_1 and I_2 and thus the epipole $\mathbf{e}$ in image I_1 can be estimated from the set of matching pairs $(\mathbf{x}_i, \mathbf{x}_i')$. Similarly, the epipole $\mathbf{e}''$ in I_3 for the camera motion corresponding to frames I_2 and I_3 can be estimated from the set of pairs $(\mathbf{x}_i', \mathbf{x}_i'')$. Since the homography from image I_2 to I_1 is simply $\mathbf{U}^{-1}$, Eq. (2) for all point matches in those two images becomes $\mathbf{x}_i \simeq \mathbf{U}^{-1}\mathbf{x}_i' + \kappa_i \mathbf{e}$. By employing Eq. (3), κ_i can then be estimated as

$$\kappa_i = \frac{(\mathbf{U}^{-1}\mathbf{x}_i' \times \mathbf{x}_i)^T (\mathbf{x}_i \times \mathbf{e})}{||\mathbf{x}_i \times \mathbf{e}||^2}. \quad (4)$$

Consider now the two image pairs (I_2, I_1) and (I_2, I_3) formed by the three images. Since image I_2 is shared by both pairs, the relative affine structure defined when I_2 becomes the first image in Eq. (2) is insensitive to the choice of the second image (i.e. I_1 or I_3) completing the pair. Therefore, applying Eq. (2) to point matches in frames I_2 and I_3 yields

$$\mathbf{x}_i'' \simeq \mathbf{V}\mathbf{x}_i' + \kappa_i \mathbf{e}'', \quad (5)$$

where the κ_i are given by Eq. (4). In order for Eq. (5) to hold for those κ_i, the scale of $\mathbf{V}$ in it has to be compatible with that of $\mathbf{e}''$. For this reason, $\mathbf{V}$ in Eq. (5) is no longer a homogeneous 3×3 matrix but rather an ordinary, inhomogeneous one. Equation (5) is thus a vector equation linear in $\mathbf{V}$, providing three linear constraints

[1] As will soon become clear, subsequent developments are valid for any homography compatible with the epipolar geometry of frames I_1 and I_2, not just for that of the quasi-metric plane.

on the nine unknown elements of $\mathbf{V}$. Due to the presence of an arbitrary, unknown scale factor, only two of those three constraints are linearly independent. Denoting the i-th row of matrix $\mathbf{V}$ by $\mathbf{v}_i^T$, writing $\mathbf{x}_i^{'\,'} = (x_i^{'\,'}, y_i^{'\,'}, 1)^T$ and $\mathbf{e}^{'\,'} = (e_x^{'\,'},\ e_y^{'\,'},\ e_z^{'\,'})^T$, those two constraints can be explicitly expressed as [2]

$$\begin{aligned} \mathbf{v}_3^T \mathbf{x}_i^{'} x_i^{'\,'} - \mathbf{v}_1^T \mathbf{x}_i^{'} &= \kappa_i e_x^{'\,'} - \kappa_i e_z^{'\,'} x_i^{'\,'} \\ \mathbf{v}_3^T \mathbf{x}_i^{'} y_i^{'\,'} - \mathbf{v}_2^T \mathbf{x}_i^{'} &= \kappa_i e_y^{'\,'} - \kappa_i e_z^{'\,'} y_i^{'\,'}. \end{aligned} \quad (6)$$

It is important to note that Eqs. (6) do not discriminate among planar and non-planar points, therefore they do not require that the employed plane has been segmented from the rest of the scene. Concatenating the equations arising from five triplet correspondences, a matrix equation of the form $\mathbf{M}\mathbf{v} = \mathbf{b}$ results, where $\mathbf{M}$ is a 10×9 matrix and $\mathbf{v} = (\mathbf{v}_1^T,\ \mathbf{v}_2^T,\ \mathbf{v}_3^T)^T$. Omitting any row of $\mathbf{M}$, yields a linear system with 9 unknowns that may be solved using linear algebra. In the case that more than five triplet matches are available, an over-constrained system results from which $\mathbf{V}$ can be estimated in a least squares manner.

As described up to this point, the estimation of $\mathbf{V}$ is achieved with a Direct Linear Transformation (DLT) algorithm [3], ch. 3. Since DLT algorithms are not invariant to similarity transformations of the point image coordinates, the accuracy of the estimation can be improved by applying the normalization technique of [7] to matching points prior to feeding them to the DLT algorithm. Independently for each image i, this normalization corresponds to a linear transformation $\mathbf{L}_i$, consisting of a translation followed by an isotropic scaling that maps the average image point to $(1, 1, 1)^T$. Notice that in this case, the normalized versions of the homography and epipole must be employed in Eq. (4) and the homography estimate $\bar{\mathbf{V}}$ computed with DLT needs to be denormalized using ${\mathbf{L}_3}^{-1}\bar{\mathbf{V}}\mathbf{L}_2$.

To prevent errors due to false matches and mislocalization of image corners from corrupting the computed homography estimate, the set of DLT constraints are considered in a robust regression framework. The LMedS estimator is employed to iteratively sample random sets of nine constraints, recover an estimate of matrix $\mathbf{V}$ from each of them and find the estimate that is consistent with the majority of the available constraints. Finally, the set of constraints having the largest support (i.e. the inliers) is employed to recompute $\mathbf{V}$ with least squares.

Since the DLT constraints minimize an algebraic error term with no physical meaning, the estimate computed by LMedS is refined by a non-linear minimization process that involves a geometric criterion. Letting $d(\mathbf{x},\ \mathbf{y})$ represent the Euclidean distance between the inhomogeneous points represented by $\mathbf{x}$ and $\mathbf{y}$, the non-linear refinement minimizes the following sum of squared distances

$$\sum_i \left(d(\mathbf{x}_i^{'\,'}, \mathbf{V}\mathbf{x}_i^{'} + \kappa_i \mathbf{e}^{'\,'})^2 + d(\mathbf{x}_i^{'}, \mathbf{V}^{-1}\mathbf{x}_i^{'\,'} - \frac{||\mathbf{x}_i^{'\,'}||}{||\mathbf{V}\mathbf{x}_i^{'} + \kappa_i \mathbf{e}^{'\,'}||}\kappa_i \mathbf{V}^{-1}\mathbf{e}^{'\,'})^2 \right) \quad (7)$$

with respect to $\mathbf{V}$. This criterion involves the mean symmetric transfer error between actual and transferred points in the two images and is minimized by applying the Levenberg-Marquardt iterative algorithm as implemented by MINPACK's LMDER routine, initialized with the estimate provided by LMedS. To safeguard against point mismatches, the non-linear refinement is performed using only the point features that correspond to inliers of the LMedS estimate.

It has already been mentioned that the entire group of all possible homography matrices between two images lies in a subspace spanned by the 4 primitive homographies of Eq. (1). This implies that given the primitive homographies for frames I_2 and I_3, the rows $\mathbf{v}_i^T$ of matrix $\mathbf{V}$ in Eqs. (6) depend on four rather than nine parameters. Therefore, the process described above can be slightly modified to estimate the coefficients λ_i making up $\mathbf{V}$ instead of directly estimating the latter. In other words, both the linear and the non-linear estimation processes can be performed with four rather than nine unknowns. It was experimentally found that the execution time for estimating $\mathbf{V}$ using the formulation involving λ_i is by an order of magnitude shorter from that required when estimating it directly.

4. PARALLAX-BASED TENSOR ESTIMATION

In this section, $\mathbf{H}_{i,j}$ and $\mathbf{e}_{i,j}$ will be used to denote respectively the virtual plane homography and the epipole in I_j for the image pair I_i and I_j. Assume also that $\mathbf{H}_{2,1}$ has been supplied and, using the method outlined in section 3, the plane homography $\mathbf{H}_{2,3}$ has been estimated from the matching triplets among images I_1, I_2 and I_3. Recalling that these homographies are, by computation, scale compatible with the corresponding epipoles, Eq. (2) yields the image projections of a 3D point $\mathbf{X}$ as $\mathbf{x} \simeq \mathbf{H}_{2,1}\mathbf{x}^{'} + \kappa\mathbf{e}_{2,1}$ and $\mathbf{x}^{''} \simeq \mathbf{H}_{2,3}\mathbf{x}^{'} + \kappa\mathbf{e}_{2,3}$, implying that $\mathbf{X} \simeq \left[\mathbf{x}^{'T},\ \kappa\right]^T$. Therefore, a consistent set of projective camera matrices in canonical form for the three views is given by [3]:

$$\mathbf{P}_1 = [\mathbf{H}_{2,1} \mid \mathbf{e}_{2,1}],\ \ \mathbf{P}_2 = [\mathbf{I} \mid \mathbf{0}],\ \ \mathbf{P}_3 = [\mathbf{H}_{2,3} \mid \mathbf{e}_{2,3}], \quad (8)$$

where $\mathbf{I}$ denotes the 3×3 identity matrix. Since it is customary to express the camera matrices relative to the first image I_1, application of an appropriate 3D projective mapping can transform Eqs.(8) so that $\mathbf{P}_1$ becomes equal to $[\mathbf{I} \mid \mathbf{0}]$. Indeed, right multiplication of the camera matrix $[\mathbf{A} \mid \mathbf{b}]$ by the 4×4 matrix $\mathbf{Q}$ given by

$$\mathbf{Q} = \left[\begin{array}{c|c} \mathbf{A}^{-1} & -\mathbf{A}^{-1}\,\mathbf{b} \\ \mathbf{0}^T & 1 \end{array}\right] \quad (9)$$

[2] Notice that all available point matches are assumed to originate from actual image points (i.e. corners); no ideal points whose third coordinate is zero exist among them.

(a) (b) (c)

Fig. 1. The first frames from (a) the Arenberg castle sequence, (b) the basement sequence and (c) the house sequence (courtesies resp. of the University of Leuven VISICS Group, Oxford University Visual Geometry Group and INRIA MOVI Group).

makes the former equal to $[\mathbf{I} \mid \mathbf{0}]$. Therefore, to make $\mathbf{P}_1$ equal to $[\mathbf{I} \mid \mathbf{0}]$, the projection matrices in Eq. (8) should be right multiplied by the matrix given by Eq.(9) for $\mathbf{A} = \mathbf{H}_{2,1}$ and $\mathbf{b} = \mathbf{e}_{2,1}$, which, taking into account that $\mathbf{H}_{j,i}^{-1}\ \mathbf{e}_{j,i} = \mathbf{e}_{i,j}$ and $\mathbf{H}_{2,3}\ \mathbf{H}_{1,2} = \mathbf{H}_{1,3}$, yields

$$\mathbf{P}_1 = [\mathbf{I} \mid \mathbf{0}]\,,\ \ \mathbf{P}_2 = [\mathbf{H}_{1,2} \mid \mathbf{e}_{1,2}]\,,\ \ \mathbf{P}_3 = [\mathbf{H}_{1,3} \mid \mathbf{e}_{1,3}]\,. \quad (10)$$

These last equations correspond to a tensor parameterization that is not minimal, but nevertheless consistent: Given the canonical projection matrices of Eqs. (10), the corresponding trifocal tensor in matrix notation is made up from the set of matrices $\{\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3\}$, where $\mathbf{T}_k = \mathbf{H}_{1,2}^k\ \mathbf{e}_{1,3}^T - \mathbf{e}_{1,2}\ {\mathbf{H}_{1,3}^k}^T$, $k = 1 \ldots 3$ and $\mathbf{H}_{i,j}^k$ denotes the k-th column of matrix $\mathbf{H}_{i,j}$ [3].

5. EXPERIMENTAL RESULTS

This section provides experimental results regarding the performance of the proposed tensor estimation algorithm. The point features that are required as input have been extracted and matched automatically. Using those matches, the epipoles were computed by finding the kernels of the fundamental matrices estimated using [8].

The first experiment was performed on the first three frames of the well-known "castle" sequence, the first of which is shown in Fig. 1(a). The 203 corners that were matched across the three frames were employed to estimate the tensor using the proposed method. To assess the accuracy of the resulting estimate, the latter was used to transfer matching corner pairs from the first two images to the third and the related root mean square (RMS) error was calculated. This error was found to be equal to 8.43 pixels. However, since certain points correspond to mismatches and thus violate the trifocal constraints, a more appropriate error measure is given by the root median square (RMedS) error, which was found to be equal to 0.69 pixels. For comparison, the RMS and RMedS errors corresponding to the tensor computed with the aid of the XCV[3] program, were equal to 7.95 and 0.66 pixels, respectively. The running time, however, of the proposed method was 130 ms on an Intel P4@1.8 GHz, a very small fraction of XCV's execution time that was in the order of 15 sec. Both running times exclude the time required to detect and match corners between images. The second experiment employed the first three frames of the "basement" sequence, the first of which is shown in Fig. 1(b). 240 corners were matched among the three frames and the RMS and RMedS errors were 57.4 and 0.79 pixels for the proposed method and 49.6 and 0.62 for XCV. Running times were 105 ms for the proposed method and around 20 sec for XCV. The third experiment was based on the "house" sequence, the first frame of which is shown in Fig. 1(c). 258 corners were matched among the three frames and the RMS and RMedS errors were 0.77 and 0.35 for the proposed method and 0.83 and 0.39 for XCV. Running time for the proposed method was 148 ms and roughly 15 sec for XCV.

[3] XCV is distributed with the VXL libraries and includes an implementation of Oxford's VGG tensor estimator [2, 3]; see also `http://vxl.sourceforge.net`

6. CONCLUSIONS

This paper has presented a computationally efficient method for estimating the trifocal tensor corresponding to three images. Although not statistically optimal in the MLE sense, the proposed method provides results of satisfactory accuracy and is fast enough to be employed in time-sensitive vision applications. The method relies on the computation of the homographies of a virtual plane, a task involving the estimation of a quadruple of plane parameters that is achieved using a combination of linear and non-linear optimization techniques operating on sets of corner matches. Due to its low dimensionality, this estimation problem can be solved efficiently, making the proposed method amenable to a fast implementation on commodity hardware. This is in contrast to conventional tensor estimation methods, which attempt to estimate at least 18 degrees of freedom simultaneously and are therefore more computationally demanding.

7. REFERENCES

[1] R.I. Hartley, "Lines and Points in Three Views and the Trifocal Tensor," *IJCV*, vol. 22, no. 2, pp. 125–140, 1997.

[2] P.H.S. Torr and A. Zisserman, "Robust Parameterization and Computation of the Trifocal Tensor," *IVC*, vol. 15, no. 8, pp. 591–605, Aug. 1997.

[3] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2000.

[4] A. Shashua and S. Avidan, "The Rank-4 Constraint in Multiple View Geometry," in *Proc. of ECCV'96*, 1996, vol. 2, pp. 196–206.

[5] S. Avidan and A. Shashua, "Tensor Embedding of the Fundamental Matrix," in *Proc. of post-ECCV SMILE'98*, 1998, vol. Springer LNCS 1506, pp. 47–62.

[6] A. Shashua and N. Navab, "Relative Affine Structure: Canonical Model for 3D from 2D Geometry and Applications," *IEEE Trans. on PAMI*, vol. 18, no. 9, pp. 873–883, Sep. 1996.

[7] R.I. Hartley, "In Defense of the 8-Point Algorithm," *IEEE Trans. on PAMI*, vol. 19, no. 6, pp. 580–593, Jun. 1997.

[8] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong, "A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry," *AI Journal*, vol. 78, pp. 87–119, 1995.