Balanced Recovery of 3D Structure and Camera Motion from Uncalibrated Image Sequences

Bogdan Georgescu[(1)] and Peter Meer[(1,2)]

[(1)] Computer Science Department,
[(2)] Electrical and Computer Engineering Department
Rutgers University, Piscataway, NJ, 08854-8058, USA
georgesc, meer@caip.rutgers.edu

Abstract. Metric reconstruction of a scene viewed by an uncalibrated camera undergoing an unknown motion is a fundamental task in computer vision. To obtain accurate results all the methods rely on bundle adjustment, a nonlinear optimization technique which minimizes the reprojection error over the structural and camera parameters. Bundle adjustment is optimal for normally distributed measurement noise, however, its performance depends on the starting point. The initial solution is usually obtained by solving a linearized constraint through a total least squares procedure, which yields a biased estimate. We present a more balanced approach where in main computational modules of an uncalibrated reconstruction system, the initial solution is obtained from a statistically justified estimator which assures its unbiasedness. Since the quality of the new initial solution is already comparable with that of the result of bundle adjustment, the burden on the latter is drastically reduced while its reliability is significantly increased. The performance of our system was assessed for both synthetic data and standard image sequences.

1 Introduction

Reliable analysis of image sequences captured by uncalibrated cameras is arguably the most significant progress in the recent years in computer vision. As the result of the analysis a 3D representation of the scene is obtained, which then can be used to acquire 3D models, generate new viewpoints, insert and delete objects, or determine the ego-motion for visual navigation. The technology became mature enough to support successful commercial ventures, such as REALVIZ or 2D3.

We follow a feature based approach toward uncalibrated image sequence analysis, in contrast with the brightness-based direct methods which consider the information from all the pixels in the image. Given an image sequence, first salient features are extracted from each frame and tracked across frames to establish correspondences. The analysis itself is a hierarchical process starting from groups of two or three images. Four main processing modules can be distinguished (Fig. 1).

1. Projective structure recovery from the key frames.
2. Insertion of the intermediate frames through camera resectioning.
3. Alignment of the independently processed subsequences.
4. Autocalibration and metric upgrade of the global reconstruction.

A. Heyden et al. (Eds.): ECCV 2002, LNCS 2351, pp. 294–308, 2002.

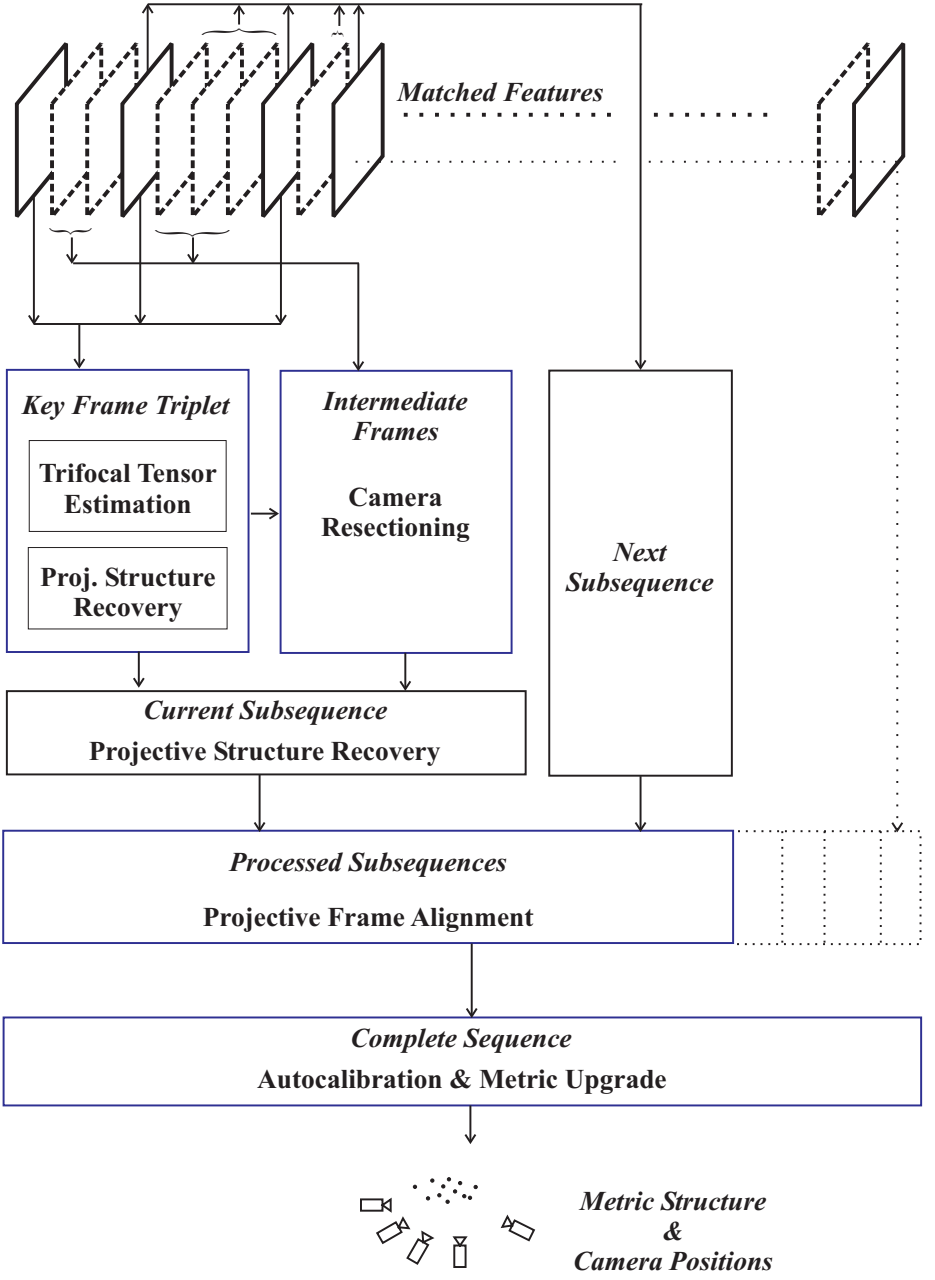

Fig. 1. The computational modules of an uncalibrated image sequence analysis system.

There is a large variety of methods proposed for each processing step. This paper will consider the most widely used techniques, described in [1], [4], [13], [14]. A nonlinear minimization problem has to be solved in each of the four modules, and most often bundle adjustment, e.g. [5], [21], is employed. Bundle adjustment is based on a sparse Levenberg-Marquardt procedure and minimizes the reprojection error over the whole set of unknown parameters, i.e., the camera matrices and the 3D structure. When the reprojection errors are normally distributed, bundle adjustment yields the optimal, maximum likelihood estimates.

The performance of any nonlinear optimization depends on the quality of the initial solution. Should this solution be too far from the true value, satisfactory convergence of the nonlinear procedure is no longer guaranteed. The traditional way to obtain the initial solution in the vision problems discussed here, is to apply a total least squares (TLS) procedure to a linearized constraint. It is well known, however, that this simple solution is biased since it fails to correctly account for the noise process that affects the *linearized* measurements, e.g. [8, p.77]. An empirical technique to improve the reliability of the linear solution is to first perform a normalizing transformation of the data [6].

In a more theoretical approach, the linearization process is analyzed and the estimation problem is put on solid theoretical foundations. The estimates obtained at the output of such methods are unbiased up to first order approximation. Kanatani's renormalization [10, pp.267–294] was the first technique from this class and was applied to a large variety of computer vision tasks. The heteroscedastic errors-in-variables (HEIV) model based estimation [11], defines the estimation problem somewhat differently than renormalization and has better numerical behavior. In this paper we show that by replacing the initial TLS based estimation step with a statistically more rigorous technique is advantageous and does not increase the amount of total computations. In fact, in some of the cases it can eliminate the need for bundle adjustment.

In Section 2 the different approaches toward solving the minimization problems arising in uncalibrated image sequence analysis are discussed. In Section 3 the performance of the four main modules are examined under two different initialization strategies: TLS and HEIV. The performance is assessed for synthetic data, while in Section 4 two standard image sequences are processed.

2 Nonlinear and Linear Minimization Techniques

Let $\boldsymbol{m}_j, j = 1, ..., n$, be the available measurements, i.e., assumed to be the unknown true values additively corrupted with normal noise having covariance $\boldsymbol{C}_{m_j}$. In the estimation process the true values are replaced with the corrected measurements $\hat{\boldsymbol{m}}_j$, and the optimal (maximum likelihood) estimates can be obtained by minimizing

$$\mathcal{J}_M = \frac{1}{2} \sum_{j=1}^{n} (\boldsymbol{m}_j - \hat{\boldsymbol{m}}_j)^\top \boldsymbol{C}_{m_j}^{+} (\boldsymbol{m}_j - \hat{\boldsymbol{m}}_j) \tag{1}$$

where $\boldsymbol{C}_{m_j}^{+}$ is the pseudoinverse.

In the most straightforward approach, the dependence of the corrected measurements $\hat{\boldsymbol{m}}_j$ on the parameter estimates $\hat{\boldsymbol{\beta}}$ is considered explicitly through a nonlinear vector

valued function, i.e., $\hat{\boldsymbol{m}}_j = \boldsymbol{f}_j(\hat{\boldsymbol{\beta}})$. The resulting unconstrained nonlinear optimization problem is called *bundle adjustment*, and it is solved using the Levenberg-Marquardt method taking also into account the sparseness of the problem [21]. For example, if the $\hat{\boldsymbol{m}}_j$-s are the measured image points corresponding to the unknown 3D points, projected with cameras whose parameters are also unknown, the criterion (1) represents the sum of squared geometric distances under the suitable Mahalanobis metric.

An alternative way of capturing the a priori geometrical information is to consider an implicit relation (constraint), between $\hat{\boldsymbol{m}}_j$ and $\hat{\boldsymbol{\beta}}$, i.e., $\boldsymbol{h}(\hat{\boldsymbol{m}}_j, \hat{\boldsymbol{\beta}}) = 0$. The minimization criterion (1) becomes

$$\mathcal{J}_M = \frac{1}{2}\sum_{j=1}^{n} (\boldsymbol{m}_j - \hat{\boldsymbol{m}}_j)^\top \boldsymbol{C}_{m_j}^{+} (\boldsymbol{m}_j - \hat{\boldsymbol{m}}_j) + \sum_{j=1}^{n} \boldsymbol{\eta}_j^\top \boldsymbol{h}(\hat{\boldsymbol{m}}_j, \hat{\boldsymbol{\beta}}) \tag{2}$$

where $\boldsymbol{\eta}_j$ are the Lagrange multipliers. In most of the problems which arise in uncalibrated image sequence analysis, this constraint can be written as

$$\boldsymbol{h}(\hat{\boldsymbol{m}}_j, \hat{\boldsymbol{\beta}}) = \boldsymbol{\Phi}(\hat{\boldsymbol{m}}_j)\,\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}, \quad j = 1, ..., n\,. \tag{3}$$

The data enters through the *carrier* matrix $\boldsymbol{\Phi}$, while the parameters are mapped through the vector valued *linearized parameter* function $\boldsymbol{\theta}$. The linear manifold structure of (3) is a consequence of the underlying projective geometry. Note that without loss of generality we can have $\|\boldsymbol{\theta}\| = 1$.

The existence of (3) motivated the use of a simple linear approximation to obtain $\hat{\boldsymbol{\theta}}$, the estimate of the linearized parameters. This estimate can then be used as initial solution for bundle adjustment. The *total least squares* (TLS) technique minimizing

$$\mathcal{J}_{TLS} = \sum_{j=1}^{n} \left\| \boldsymbol{\Phi}(\boldsymbol{m}_j)\hat{\boldsymbol{\theta}} \right\|^2 \tag{4}$$

i.e., the algebraic distance from the hyperplane with unit normal $\hat{\boldsymbol{\theta}}$, is most often employed. The TLS estimator, however, is optimal only when *all* the rows $\boldsymbol{\phi}_k$ of the carrier matrix are corrupted with the same noise process, which must have covariance $\sigma^2\boldsymbol{I}$ [22, Sec. 8.2]. This is not true for the estimation problems under consideration even when all $\boldsymbol{C}_{m_j} = \sigma^2\boldsymbol{I}$, since the elements of the carrier matrix are nonlinear functions in the measurements $\boldsymbol{m}_j$.

Analyzing the structure of the carrier matrix $\boldsymbol{\Phi}$ reveals that the noise process which has to be considered when (2) is minimized, is point dependent, i.e., heteroscedastic. The *heteroscedastic errors-in-variables* (HEIV) estimator described in [11] takes into account the nature of the noise process and finds $\hat{\boldsymbol{\theta}}$ by solving iteratively the generalized eigenproblem

$$\nabla_{\hat{\boldsymbol{\theta}}} \mathcal{J}_M = [\boldsymbol{S}(\hat{\boldsymbol{\theta}}) - \boldsymbol{C}(\hat{\boldsymbol{\theta}})]\hat{\boldsymbol{\theta}} = \boldsymbol{0} \quad \text{subject to} \quad \|\hat{\boldsymbol{\theta}}\| = 1 \tag{5}$$

where

$$\boldsymbol{S}(\hat{\boldsymbol{\theta}}) = \sum_{j=1}^{n} \boldsymbol{\Phi}(\boldsymbol{m}_j)^\top \hat{\boldsymbol{\Sigma}}_j^{+} \boldsymbol{\Phi}(\boldsymbol{m}_j) \qquad \hat{\boldsymbol{\Sigma}}_j = \hat{\boldsymbol{\theta}}^\top \boldsymbol{J}_{\Phi|\hat{m}_j}^\top \boldsymbol{C}_{m_j} \boldsymbol{J}_{\Phi|\hat{m}_j} \hat{\boldsymbol{\theta}} \tag{6}$$

is the scatter matrix. The Jacobian matrices $\boldsymbol{J}_{\boldsymbol{\Phi}|\hat{\boldsymbol{m}}_j} = \partial\boldsymbol{\Phi}(\hat{\boldsymbol{m}}_j)/\partial\hat{\boldsymbol{m}}_j$ can be easily computed since most of the elements of the carrier matrix have a multilinear structure. The expression of the weighted covariance matrix is

$$\boldsymbol{C}(\hat{\boldsymbol{\theta}}) = \sum_{j=1}^{n}\sum_{k,l}\eta_{kj}\eta_{lj}\left[\frac{\partial\phi_k(\hat{\boldsymbol{m}}_j)}{\partial\hat{\boldsymbol{m}}_j}\right]^{\top}\boldsymbol{C}_{m_j}\left[\frac{\partial\phi_l(\hat{\boldsymbol{m}}_j)}{\partial\hat{\boldsymbol{m}}_j}\right] \qquad \boldsymbol{\eta}_j = \hat{\boldsymbol{\Sigma}}_j^{+}\boldsymbol{\Phi}(\boldsymbol{m}_j)\hat{\boldsymbol{\theta}} \tag{7}$$

where ϕ_k is the k^{th} row of $\boldsymbol{\Phi}$. The corrected measurements are

$$\hat{\boldsymbol{m}}_j = \boldsymbol{m}_j - \boldsymbol{C}_{m_j}\boldsymbol{J}_{\boldsymbol{\Phi}|\hat{\boldsymbol{m}}_j}\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\Sigma}}_j^{+}\boldsymbol{\Phi}(\boldsymbol{m}_j)\hat{\boldsymbol{\theta}} \tag{8}$$

and analytical expressions for the covariances of the estimated parameters $\boldsymbol{C}_{\hat{\theta}}$ and the corrected measurements $\boldsymbol{C}_{\hat{m}_j}$ are also available [11, Sec. 5.2].

In spite of being computed iteratively, the performance of the HEIV estimation is not critically dependent of the initial choice of $\hat{\boldsymbol{\theta}}$ in (5). Indeed, in most of the cases using a random initial $\hat{\boldsymbol{\theta}}$ suffices. A more accurate starting value can be obtained by approximating the initial $\boldsymbol{S}(\hat{\boldsymbol{\theta}})$ and $\boldsymbol{C}(\hat{\boldsymbol{\theta}})$ from the available measurements $\boldsymbol{m}_j$ and their covariances $\boldsymbol{C}_{m_j}$ (usually taken as $\sigma^2\boldsymbol{I}$) [11, Sec. 5.6]. After each iteration the measurements are corrected (8) and the Jacobian matrices are updated followed by $\boldsymbol{S}(\hat{\boldsymbol{\theta}})$ and $\boldsymbol{C}(\hat{\boldsymbol{\theta}})$. Convergence is usually reached after 3-4 iterations. It can be shown that $\hat{\boldsymbol{\theta}}$ is an unbiased estimate at the first order approximation [11, Sec. 5.2].

The minimization criterion solved by the HEIV estimator is similar to that of the Sampson distance [8, Sec. 15.4.3]. However, traditionally when the Sampson distance is used, the solution is still obtained through the Levenberg Marquardt algorithm and the Jacobian matrices are not updated at each step [8, pp. 387–388].

The parameter of interest in the optimization is $\boldsymbol{\beta}$ and not $\boldsymbol{\theta}$. Since at each iteration of the HEIV estimation procedure the covariance of the current linearized parameter estimates $\boldsymbol{C}_{\hat{\theta}}$ is available, $\hat{\boldsymbol{\theta}}$ can be further refined by imposing the constraint of its nonlinear dependence on $\boldsymbol{\beta}$ (3). This is achieved by projecting $\hat{\boldsymbol{\theta}}$ under the metric induced by $\boldsymbol{C}_{\hat{\theta}}$ on the nonlinear manifold in the space of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\boldsymbol{\beta})\right\|^2_{\boldsymbol{C}_{\hat{\theta}}} \tag{9}$$

which is solved by linearization [11, Sec. 5.10]. The parameter estimate $\hat{\boldsymbol{\theta}}$ can now be updated as $\hat{\boldsymbol{\theta}}^{(u)} = \boldsymbol{\theta}(\hat{\boldsymbol{\beta}})$ and this value is used in the next iteration of the HEIV estimator.

3 Uncalibrated Image Sequence Analysis System

The reconstruction of the 3D scene in the system described in this paper is based on point features detected in the images. The Harris corner detector was used, since it provides the most stable features under a wide range of operating conditions [16]. The correspondences across frames are established by the traditional normalized cross-correlation technique.

The first step in the analysis is to break down the sequence into several small groups of *key frames*. Since the "local" estimation of the projective structure employs the trifocal tensor, each group has three key frames. Given the first frame, the second and third frames are chosen to satisfy the trade-off between increasing the baseline of the group and having enough reliably tracked features. Adjacent key frames triplets have two frames in common (Fig. 1).

The three key frames delimit sets of contiguous *intermediate frames* in the sequence. Because the salient points were tracked also through these frames, the projection matrices for each intermediate frame can be computed by camera resectioning. The projective structure is then refined for the entire subsequence. The same process is applied independently to the next triplet of key frames, and the two subsequences are aligned by bringing them into the same projective basis.

After the entire available image sequence was processed and aligned, given that the camera motion is not degenerate, the metric structure of the 3D scene is recovered by imposing additional constraints on the internal camera parameters.

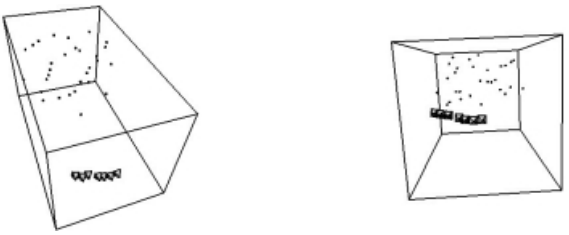

Fig. 2. Synthetic data. Two views of the typical configuration.

The role of the two initialization methods in the four computational modules will be assessed using synthetic data. Thirty 3D points were uniformly distributed in a cube and kept in the field of view in each of the 7 frames of size 512x512. The projected points were corrupted by normal noise with standard deviation $\sigma = 0.5$ pixel units. Every second frame was taken as a key frame, thus having two subsequences of five frames each. The performance of the four modules was recorded in 100 trials. Between the trials the measurement noise is changed, and the position of the cameras was slightly perturbed by a random displacement. Two views of a typical experimental configuration is shown is Fig. 2. Note the small baseline of the camera movement which increases the difficulty of the processing. To assess the performance of an individual module, the output of bundle adjustment from the previous module was used.

3.1 Projective Structure Recovery from the Key Frames

The first computational module of the uncalibrated image sequence analysis system recovers the projective structure defined by triplets of key frames. The employed geometric constraint is based on the trifocal tensor which describes, independently of the scene structure, the intrinsic properties of the group of three images [17].

The incidence relation between the three point projections $\{\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{x}''\}$ corresponding to the same 3D point can be written using the estimated $3 \times 3 \times 3$ trifocal tensor $\boldsymbol{T}$ as

$$[\boldsymbol{x}']_{\times} \left(\sum_{i=1}^{3} x_i \boldsymbol{T}_i \right) [\boldsymbol{x}'']_{\times} = \boldsymbol{0}_{3\times 3} \tag{10}$$

where $[\boldsymbol{v}]_{\times}$ is the skew-symmetric matrix such that $\boldsymbol{v} \times \boldsymbol{u} = [\boldsymbol{v}]_{\times} \boldsymbol{u}$ and the 3×3 matrices $\boldsymbol{T}_i$ are the correlation slices of the trifocal tensor [2]. The trifocal tensor is related to the projection matrices of the three frames $\boldsymbol{P} = [\boldsymbol{I}|\boldsymbol{0}]$, $\boldsymbol{P}' = [\boldsymbol{A}|\boldsymbol{e}']$ and $\boldsymbol{P}'' = [\boldsymbol{B}|\boldsymbol{e}'']$ by

$$\boldsymbol{T}_i = \boldsymbol{a}_i \boldsymbol{e}''^{\top} - \boldsymbol{e}' \boldsymbol{b}_i^{\top} . \tag{11}$$

The constraint is satisfied by the *true* values of the quantities involved. For the estimation the relation (10) can be easily rewritten under the from (3)

$$\boldsymbol{\Phi}_t(\hat{\boldsymbol{m}}_t)\boldsymbol{\theta}_t(\hat{\boldsymbol{\beta}}_t) = \boldsymbol{0}_9 \qquad \|\boldsymbol{\theta}_t\| = 1 \tag{12}$$

where the elements of the carrier matrix are products of the corrected image coordinates $\hat{\boldsymbol{m}}_t = [\hat{x}_1, \hat{x}_2, \hat{x}'_1, \hat{x}'_2, \hat{x}''_1, \hat{x}''_2]^{\top} \in \mathbb{R}^6$ and the components of the unconstrained parameter vector $\boldsymbol{\theta}_t \in \mathbb{R}^{27}$ are the estimates of the trifocal tensor elements. Each point correspondence contributes with 4 independent constraints [8, pp.417–421].

It can be shown that the trifocal tensor has only 18 degrees of freedom [2], and different parametrizations can be employed to constrain the 27 values of $\boldsymbol{\theta}_t$ to represent a tensor [18]. We have used the 24 parameters of the projection matrices $\boldsymbol{P}'$ and $\boldsymbol{P}''$ for parametrization, thus $\boldsymbol{\beta}_t \in \mathbb{R}^{24}$. Note that only 22 parameters are significant due to the scale ambiguity of the projection matrices.

Since not all point correspondences are correct, the estimation process must be implemented robustly. Instead of the traditional RANSAC approach we have used one of its variants MLESAC [19], and minimized the transfer error, i.e., the robust sum of squared distances between the measurements and the corrected points. Subsequent computations were based only on the inliers. We have found that when the percentage of erroneous matches is small (say under 20%) a global M-estimation procedure is already satisfactory. This condition can be assured by using a high correlation score threshold.

The optimal, Gold Standard method for the recovery of the projective structure from a triplet of the key frames, is to apply bundle adjustment over the camera parameters and the 3D position of each feature [3], [18]. The initial solution is computed by recovering the camera matrices from the tensor (11), and using this information to obtain the 3D coordinates of each point by triangulation. The initial solution was computed with either TLS using normalized image coordinates [6] or HEIV. In the latter case, the corrected measurements $\hat{\boldsymbol{m}}_t$ are also available (8) and the camera parameters are obtained using the estimation process (9).

The performance was assessed through the reprojection error, i.e., the root-mean-squared (RMS) residual error which is proportional to the square root of the optimization criterion value (1). In Fig. 3 the histograms of the reprojection errors are shown for the different processing methods. As expected, the linear TLS solution is of poor quality being strongly biased (Fig. 3a), though bundle adjustment succeeds to eliminate this bias (Fig. 3b). The HEIV solution (Fig. 3c), on the other hand, already returns the same

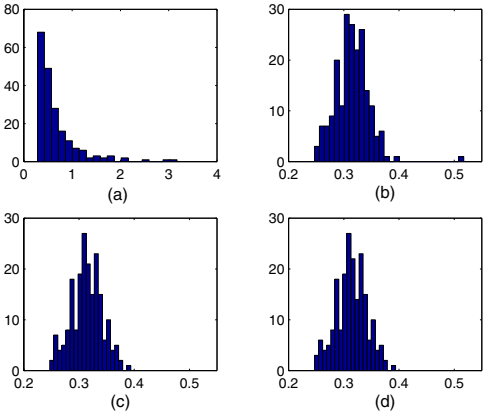

Fig. 3. Reprojection errors for the trifocal tensor estimation from the key frames. (a) TLS initial solution. Note the different scale from HEIV. (b) Bundle adjustment initiated with TLS. (c) HEIV initial solution. (d) Bundle adjustment initiated with HEIV.

estimate as the Gold Standard method, subsequent bundle adjustment is not necessary since it will not yield any improvement (Fig. 3d). The number of bundle adjustment iterations for the TLS initialization was on average 5.3 but with high variation ($\sigma_{it} = 6.5$) while using the HEIV initialization no additional iterations were needed. See [12] for a detailed discussion about using HEIV method for the trifocal tensor estimation.

3.2 Insertion of the Intermediate Frames

To complete the projective structure estimation for the entire subsequence defined by the three selected key frames, the information provided by the intermediate frames must be also integrated. Will denote with $\boldsymbol{X}$ the projective coordinates of the 3D points, whose image was tracked *through* all the intermediate frames. From the processing of the key frames, the estimates of these 3D points are already available. Thus, using camera resectioning [8, pp. 166–170] the initial solution for the camera matrices of the intermediate frames can be determined.

The projective image formation relation $\boldsymbol{x} \sim \boldsymbol{P}\boldsymbol{X}$ can be rewritten as

$$\begin{bmatrix} \mathbf{0}_4^\top & -x_3\boldsymbol{X}^\top & x_2\boldsymbol{X}^\top \\ x_3\boldsymbol{X}^\top & \mathbf{0}_4^\top & -x_1\boldsymbol{X}^\top \\ -x_2\boldsymbol{X}^\top & x_1\boldsymbol{X}^\top & \mathbf{0}_4^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{p}^{(1)} \\ \boldsymbol{p}^{(2)} \\ \boldsymbol{p}^{(3)} \end{bmatrix} = \mathbf{0}_3 \tag{13}$$

where we denote by $\boldsymbol{p}^{(k)\top}$ the k^{th} row of the camera matrix $\boldsymbol{P}$. Each measurement contributes with two linear independent equations, and thus the images of at least six 3D points must be available. From the constraint (13), for the estimation we obtain an expression which has the form (3)

$$\boldsymbol{\Phi}_r\left(\hat{\boldsymbol{m}}_r\right)\hat{\boldsymbol{\theta}}_r = \mathbf{0}_3 \tag{14}$$

where the carrier matrix $\boldsymbol{\Phi}_r$ has as elements products between the corrected projective image and 3D coordinates, $\hat{\boldsymbol{m}}_r = [\hat{\boldsymbol{x}}^\top, \hat{\boldsymbol{X}}^\top]^\top$, and $\hat{\boldsymbol{\theta}}_r = \mathrm{vec}[\hat{\boldsymbol{P}}^\top]$ are the elements of the projection matrix to be estimated.

The projection matrices are estimated for each of the intermediate frames initially by TLS and HEIV. Note that the HEIV based estimation takes into account that the "measurements" of the 3D points $\boldsymbol{X}_i$ (estimated in the first module) have covariances $\boldsymbol{C}_{X_i}$. The entire subsequence, defined by the key frames and the intermediate frames, is passed to the bundle adjustment which refines globally the camera parameters and the 3D points coordinates.

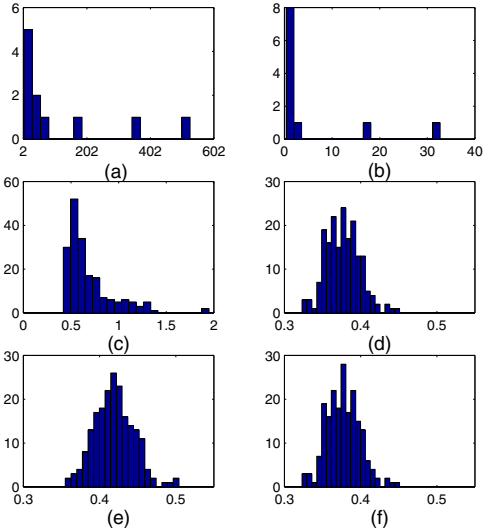

Fig. 4. Reprojection errors for the entire subsequence. (a) TLS initial solutions, large errors. (b) Bundle adjustment initiated with the estimates in (a). (c) TLS initial solution. Note the different scale from HEIV. (d) Bundle adjustment initiated with TLS. (e) HEIV initial solution. (f) Bundle adjustment initiated with HEIV.

Results obtained with TLS initialization are presented in Figs. 4a and 4c, while the output of the corresponding bundle adjustment is shown in Figs. 4b and 4d. A few of the TLS initializations fail yielding RMS errors larger that 2 (Fig. 4a). Bundle adjustment did not succeed to recover from all of these cases (Fig. 4b). The bias of the TLS initial solution is also visible in Fig. 4c but it is removed after bundle adjustment (Fig. 4d). The HEIV initial solution is unbiased and yields smaller errors than the TLS (Fig. 4e). Since bundle adjustment is a global procedure, the errors are further reduced (Fig. 4f). It should be emphasized that after bundle adjustment both initializations give the same results (except the few failures of TLS) but fewer iterations were needed for bundle adjustment to converge for the HEIV initialization (average 3.5 with $\sigma_{it} = 0.8$) than using the TLS initialization (average 5.1 but with large $\sigma_{it} = 6.1$).

In the same framework we can also approach the triangulation procedure, i.e. finding the location of the 3D projective coordinates of a point knowing its projection in several images *and* the camera matrices. While it was not used in the performance comparisons with synthetic data, triangulation is an important step in the analysis of real image

sequences to obtain additional point correspondences and augment the available structure [7].

For triangulation, in (13) the parameters become the 3D projective point coordinates, and the "measurements" are the image points and camera matrices

$$\boldsymbol{\Phi}_g(\boldsymbol{m}_g)\boldsymbol{\theta}_g = \begin{bmatrix} x_1\boldsymbol{p}^{(3)\top} - x_3\boldsymbol{p}^{(1)\top} \\ x_2\boldsymbol{p}^{(3)\top} - x_3\boldsymbol{p}^{(2)\top} \end{bmatrix} \boldsymbol{X} = \mathbf{0}_2\,. \tag{15}$$

The HEIV estimation takes into account the nonliniarities present in the carrier matrix, i.e., the products of image coordinates and camera matrix elements.

3.3 Alignment of the Independently Processed Subsequences

After a subsequence was processed, the newly obtained structure must be aligned with the already recovered structure. This can be achieved since there are at least two frames overlapping with the previous subsequence.

Assume that the frame j is shared by both subsequences and an image point $\boldsymbol{x}_{ij}$ from this frame corresponds to the 3D point having the coordinates $\boldsymbol{X}_i$ and $\boldsymbol{X}'_i$ in the projective base of the two subsequences. Then the homography $\boldsymbol{H}$ that aligns the subsequences must obey

$$\boldsymbol{x}_{ij} \sim \boldsymbol{P}_j\boldsymbol{X}_i \sim \boldsymbol{P}'_j\boldsymbol{H}\boldsymbol{H}^{-1}\boldsymbol{X}'_i \quad \text{or} \quad \boldsymbol{P}_j \sim \boldsymbol{P}'_j\boldsymbol{H} \quad \boldsymbol{X}_i \sim \boldsymbol{H}^{-1}\boldsymbol{X}'_i \tag{16}$$

where $\boldsymbol{P}_j$ and $\boldsymbol{P}'_j$ are the projective matrices of the frame j in the two bases.

Different methods allowing linear solutions for $\boldsymbol{H}$, based on direct 3D point registration, which is not meaningful in a projective framework, enforcing camera consistency, or a combination of these two were proposed [4]. We use the reprojection error between $\boldsymbol{P}'_j\boldsymbol{H}\boldsymbol{X}_i$ and the corresponding image coordinates $\boldsymbol{x}_{ij}$. From (16) this constraint is

$$\boldsymbol{x}_{ij} \sim \boldsymbol{P}'_j\boldsymbol{H}\boldsymbol{X}_i \tag{17}$$

which can be expressed for estimation as

$$\boldsymbol{\Phi}_h\left(\hat{\boldsymbol{m}}_h\right)\hat{\boldsymbol{\theta}}_h = \mathbf{0}_2 \tag{18}$$

where the elements of the carrier matrix $\boldsymbol{\Phi}_h$ are triple products of the corrected image coordinates, projection matrix elements and projective coordinates of the 3D points, $\hat{\boldsymbol{m}}_h^\top = [\hat{\boldsymbol{x}}^\top, \text{vec}[\hat{\boldsymbol{P}}^\top]^\top, \hat{\boldsymbol{X}}^\top]$ and $\hat{\boldsymbol{\theta}}_h$ contains the 16 components of the homography $\boldsymbol{H}$ to be estimated.

The TLS initialization started from the results of the TLS based bundle adjustment from the previous module. Thus the few cases yielding large residual errors were also considered. Some of these cases were successfully processed by alignment, however, other new ones were introduced (Figs. 5a and 5b). The HEIV initialization used the estimated covariance matrices $\boldsymbol{C}_{\hat{P}'}$ and $\boldsymbol{C}_{\hat{X}}$ of the projection matrices and 3D points. No failures were obtained and its performance (Fig. 5e) is further refined by the global bundle adjustment (Fig. 5f). The average number of iterations using the TLS initialization was 6.4 with $\sigma_{it} = 9.42$ while using the HEIV initialization the average was 3.9 with $\sigma_{it} = 1.6$.

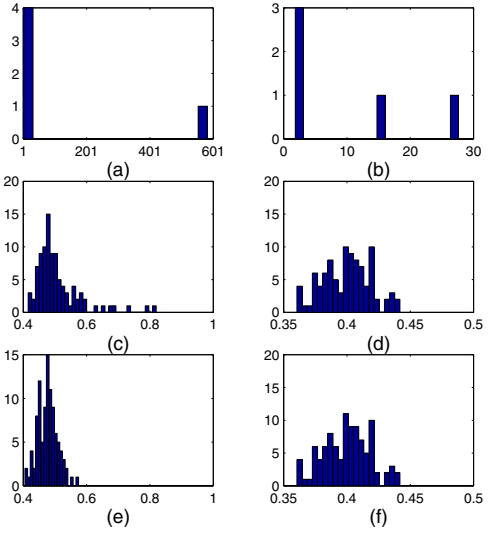

Fig. 5. Reprojection errors for alignment of subsequences. (a) TLS initial solutions, large errors. (b) Bundle adjustment initiated with the estimates in (a). (c) TLS initial solution. (d) Bundle adjustment initiated with TLS. (e) HEIV initial solution. (f) Bundle adjustment initiated with HEIV.

3.4 Autocalibration and Metric Upgrade of the Global Reconstruction

The autocalibration method used in this paper is based on the dual absolute quadric $\boldsymbol{\Omega}^*$ and its relation to the dual image of the absolute conic $\boldsymbol{\omega}_j^*$ [20]

$$\boldsymbol{\omega}_j^* \sim \boldsymbol{K}_j \boldsymbol{K}_j^\top \sim \boldsymbol{P}_j \boldsymbol{\Omega}^* \boldsymbol{P}_j^\top \tag{19}$$

where $\boldsymbol{K}_j$ are the internal camera parameters for frame j. When additional knowledge about the internal parameters is available, such as no skew, known principal point or aspect ratio, then relation (19) can be used to obtain constraints on the dual absolute quadric [9], [14], [15]. If we assume that the aspect ratio is one, the skew is zero and the principal point is in the center of the image, then (19) yields four linear independent equations

$$\begin{aligned} \boldsymbol{p}_j^{(1)\top} \boldsymbol{\Omega}^* \boldsymbol{p}_j^{(1)} &= \boldsymbol{p}_j^{(2)\top} \boldsymbol{\Omega}^* \boldsymbol{p}_j^{(2)} \\ \boldsymbol{p}_j^{(i)\top} \boldsymbol{\Omega}^* \boldsymbol{p}_j^{(k)} &= 0 \qquad (i,k) \in \{(1,2),(1,3),(2,3)\} \end{aligned} \tag{20}$$

which can be rearranged for estimation as in (3)

$$\boldsymbol{\Phi}_a(\hat{\boldsymbol{m}}_a)\hat{\boldsymbol{\theta}}_a = \mathbf{0}_4 \tag{21}$$

where the carriers $\boldsymbol{\Phi}$ have as elements double products of the projection matrix elements, $\hat{\boldsymbol{m}}_a = vec[\hat{\boldsymbol{P}}^\top] \in \mathbb{R}^{12}$ and $\hat{\boldsymbol{\theta}}_a \in \mathbb{R}^{10}$ contains the dual absolute quadric elements to be estimated. Because of the symmetry only 10 such elements are needed.

If one of the projection matrices is chosen as reference $\boldsymbol{P}_0 = [\boldsymbol{I}|\boldsymbol{0}_3]$ then $\boldsymbol{\Omega}^*$ becomes

$$\boldsymbol{\Omega}^* = \begin{bmatrix} \boldsymbol{K}_0\boldsymbol{K}_0^\top & -\boldsymbol{K}_0\boldsymbol{K}_0^\top\boldsymbol{\pi}_\infty \\ -\boldsymbol{\pi}_\infty^\top\boldsymbol{K}_0\boldsymbol{K}_0^\top & -\boldsymbol{\pi}_\infty^\top\boldsymbol{K}_0\boldsymbol{K}_0^\top\boldsymbol{\pi}_\infty \end{bmatrix} \tag{22}$$

where $\boldsymbol{\pi}_\infty$ defines the plane at infinity. Thus $\boldsymbol{\Omega}^*$ can be parametrized by maximum 8 parameters (3 from the plane at infinity and the rest from $\boldsymbol{K}_0$). The following transformation brings the recovered projective structure into a metric reconstruction

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{K}_0 & \boldsymbol{0} \\ -\boldsymbol{\pi}_\infty\boldsymbol{K}_0^\top & 1 \end{bmatrix}. \tag{23}$$

Following [15] the TLS initial solution was further refined by solving with Levenberg-Marquardt the nonlinear least squares problem

$$\sum_j \left\| \frac{\boldsymbol{K}_j\boldsymbol{K}_j^\top}{\|\boldsymbol{K}_j\boldsymbol{K}_j^\top\|_F} - \frac{\boldsymbol{P}_j\boldsymbol{\Omega}^*\boldsymbol{P}_j^\top}{\|\boldsymbol{P}_j\boldsymbol{\Omega}^*\boldsymbol{P}_j^\top\|_F} \right\|_F^2 \tag{24}$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix. The refinement of HEIV initialization was based on the nonlinear correction (9) which provided the estimates of the parametrization (22).

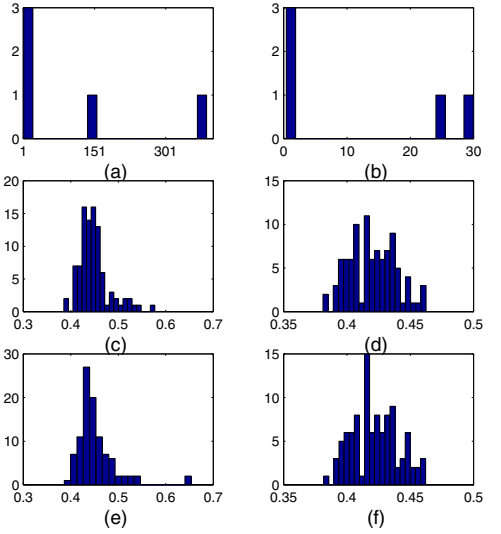

Fig. 6. Reprojection errors for metric upgrade of the entire structure. (a) Refined TLS initial solutions, large errors. (b) Bundle adjustment initiated with the estimates in (a). (c) Refined TLS initial solution. (d) Bundle adjustment initiated with refined TLS. (e) HEIV initial solution. (f) Bundle adjustment initiated with HEIV.

The projective reconstruction is upgraded to a metric reconstruction using the homography computed from (23) and bundle adjustment is employed over all the available 3D

points, internal and external camera parameters. The rotation matrices were parametrized with quaternions.

The results before and after metric bundle adjustment are presented in Fig. 6. The HEIV based initial solution (Fig. 6e) has similar performance to the combined TLS and nonlinear LS solution for the majority of the data (Fig. 6c) while not having the spurious large residual errors (Fig. 6a).

4 Experimental Results with Image Sequences

The system using the HEIV initialization was run on two well known real image sequences. Fig. 7 shows two images from the *MOVI house* image sequence and two poses of the reconstructed scene and camera positions. The sequence has 118 frames of a scene taken by moving the objects on a turntable. Significant illumination changes appear in the sequence because the objects were moved with respect to the light source. Note also that the density of the frames is not uniform. The reconstruction was computed automatically and without imposing constraints on the camera motion. After metric upgrade, the cameras that were close in 3D were used to establish additional correspondences which helped to improve the alignment of the entire sequence. It can be seen that the camera positions are lying on a planar circular path while keeping the scene in the field of view. The reconstructed position in 3D space of the scene features obey the rectangular shape of the house and lay on circular surfaces for the can and cup.

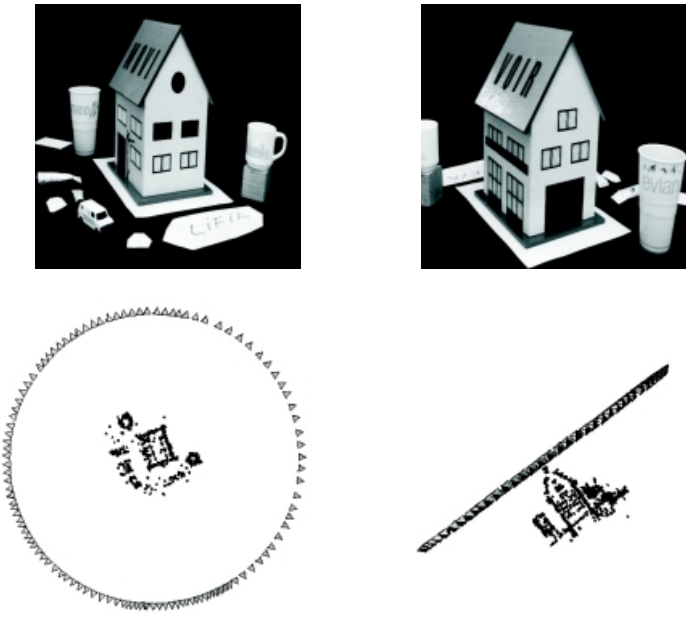

Fig. 7. Metric reconstruction of the MOVI house sequence.

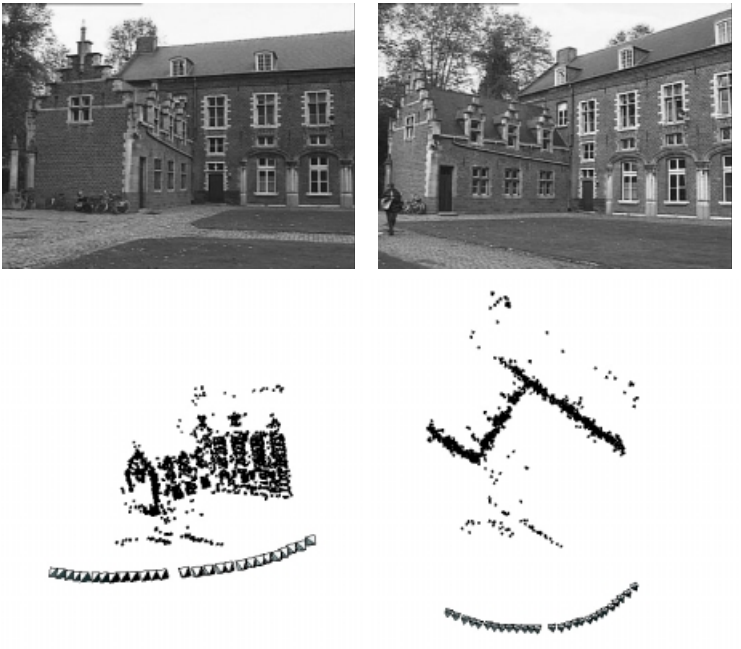

Fig. 8. Metric reconstruction of the castle sequence

The second image sequence processed was the *castle* sequence. In Fig. 8 two images from the sequence and the reconstructed structure are shown. The sequence is 27 frames long and contains also some small nonrigid elements. The metric reconstruction successfully recovers the main features of the scene

5 Conclusion

We have presented a detailed investigation of the importance of using a statistically accurate initialization procedure in all the processing modules of an uncalibrated image sequence analysis system. The reliability of the system is further increased, and the failures for difficult data may be avoided.

Acknowledgment. The support of the NSF grant IRI 99-87695 is gratefully acknowledged. We thank Dr. Bogdan Matei from Sarnoff Corporation for insightful discussions.

References

1. P. Beardsley, P. Torr, and A. Zisserman. 3D model aquisition from extended image sequences. In B. Buxton and R. Cipolla, editors, *Computer Vision – ECCV 1996*, volume II, pages 683–695, Cambridge, UK, April 1996. Springer.
2. N. Canterakis. A minimal set of constraints for the trifocal tensor. In D. Vernon, editor, *Computer Vision – ECCV 2000*, volume I, pages 84–99, Dublin, Ireland, 2000. Springer.

3. O. Faugeras and T. Papadopoulo. A nonlinear method for estimating the projective geometry of 3 views. In *6th International Conference on Computer Vision*, pages 477–484, Bombay, India, January 1998.
4. A.W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. In H. Burkhardt and B. Neumann, editors, *Computer Vision – ECCV 1998*, volume I, pages 311–326, Freiburg, Germany, June 1998. Springer.
5. R. I. Hartley. Euclidean reconstruction from uncalibrated views. In J.L. Mundy, A. Zisserman, and D. Forsyth, editors, *Applications of Invariance in Computer Vision*, pages 237–256, 1994.
6. R. I. Hartley. In defence of the 8-point algorithm. In *5th International Conference on Computer Vision*, pages 1064–1070, Cambridge, MA, June 1995.
7. R. I. Hartley and P. Sturm. Triangulation. *Computer Vision and Image Understanding*, 68:146–157, 1997.
8. R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.
9. A. Heyden and K. Astrom. Euclidean reconstruction from image sequences with varying and unknown focal length and principal point. In *1997 IEEE Conference on Computer Vision and Pattern Recognition*, pages 438–443, San Juan, Puerto Rico, June 1997.
10. K. Kanatani. *Statistical Optimization for Geometric Computation: Theory and Practice*. Elsevier, 1996.
11. B. Matei. *Heteroscedastic Errors-In-Variables Models in Computer Vision*. PhD thesis, Department of Electrical and Computer Engineering, Rutgers University, 2001. Available at `http://www.caip.rutgers.edu/riul/research/theses.html`.
12. B. Matei, B. Georgescu, and P. Meer. A versatile method for trifocal tensor estimation. In *8th International Conference on Computer Vision*, volume II, pages 578–585, Vancouver, Canada, July 2001.
13. P.F. McLauchlan and D.W. Murray. A unifying framework for structure and motion recovery from image sequences. In *5th International Conference on Computer Vision*, pages 314–320, Cambridge, Massachusetts, June 1995.
14. M. Pollefeys. *Self-calibration and Metric 3D Reconstruction from Uncalibrated Image Sequences*. PhD thesis, K. U. Leuven, 1999.
15. M. Pollefeys. Self calibration and metric reconstruction in spite of varying and unknown intrinsic camera parameters. *International J. of Computer Vision*, 32:7–25, 1999.
16. C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *Computer Vision and Image Understanding*, 78:151–172, 2000.
17. A. Shashua. Algebraic functions for recognition. *IEEE Trans. Pattern Anal. Machine Intell.*, 17:779–780, 1995.
18. P. H. S. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing*, 15:591–605, August 1997.
19. P.H.S. Torr and A. Zisserman. Robust computation and parametrization of multiple view relations. In *6th International Conference on Computer Vision*, pages 727–732, Bombay, India, January 1998.
20. B. Triggs. Autocalibration and the absolute quadric. In *1997 IEEE Conference on Computer Vision and Pattern Recognition*, pages 609–614, San Juan, Puerto Rico, June 1997.
21. B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustment — A modern synthesis. In B. Triggs, A. Zisserman, and R. Szelisky, editors, *Vision Algorithms: Theory and Practice*, pages 298–372. Springer, 2000.
22. S. Van Huffel and J. Vanderwalle. Analysis and properties of GTLS in problem $AX \approx B$. *SIAM Journal on Matrix Analysis and Applications*, 10:294–315, 1989.