Real-time Quadrifocal Visual Odometry

A. I. Comport
CNRS Laboratoire I3S,
2000 route des Lucioles,
Sophia-Antipolis,
France
comport@i3s.unice.fr

E. Malis
P. Rives
INRIA, Sophie-Antipolis Mediterrane
2004 route des Lucioles,
Sophia-Antipolis,
France
ezio.malis@sophia.inria.fr
Patrick.Rives@sophia.inria.fr

Abstract

In this paper we describe a new image-based approach to tracking the six-degree-of-freedom trajectory of a stereo camera pair. The proposed technique estimates the pose and subsequently the dense pixel matching between temporal *image pairs in a sequence by performing dense* spatial *matching between images of a stereo reference pair. In this way a minimization approach is employed which directly uses all grayscale information available within the stereo pair (or stereo region) leading to very robust and precise results. Metric 3D structure constraints are imposed by consistently warping corresponding stereo images to generate novel viewpoints at each stereo acquisition. An iterative non-linear trajectory estimation approach is formulated based on a quadrifocal relationship between the image intensities within adjacent views of the stereo pair. A robust M-estimation technique is used to reject outliers corresponding to moving objects within the scene or other outliers such as occlusions and illumination changes. The technique is applied to recovering the trajectory of a moving vehicle in long and difficult sequences of images.*

KEY WORDS—visual odometry, direct, dense 3D tracking stereo, multi-view geometry

The International Journal of Robotics Research
Vol. 29, No. 2–3, February/March 2010, pp. 245–266
DOI: 10.1177/0278364909356601

Figures 1–3, 7–9, 12, 13 appear in color online: http://ijr.sagepub.com

1. Introduction

Here the core issue of 3D visual odometry is considered in the context of rapidly moving vehicles, real sequences, large-scale distances, with traffic and other types of occluding information. Indeed, tracking in urban canyons is a non-trivial problem (Mouragnon et al. 2006; Simond and Rives 2008). It is clear that pose estimation and visual tracking are also important in many applications including robotic vision, augmented reality, medical imaging, etc.

Model-based techniques have shown that 3D CAD models are essential for robust, accurate and efficient 3D motion estimation (Comport et al. 2006b), however, they have the major drawback of requiring an *a priori* model which is not always available or extremely difficult to obtain as in the case of shapeless objects or large-scale environments.

Alternative techniques propose to perform 3D structure and motion estimation online. Among this class, visual simultaneous localization and mapping approaches (Chiuso et al. 2002; Davison and Murray 2002) are classically based on an implementation of the extended Kalman filter and have limited computational efficiency (manipulation and inversion of large feature co-variance matrices) and limited inter-frame movements (owing to approximate non-iterative estimation). Nistér et al. (2004) proposed stereo and monocular visual odometry approaches based on a combination of feature extraction, matching, tracking, triangulation, RANSAC pose estimation and iterative refinement. Mouragnon et al. (2006) proposed a similar monocular technique, but in this approach drift is minimized using a local bundle adjustment technique.

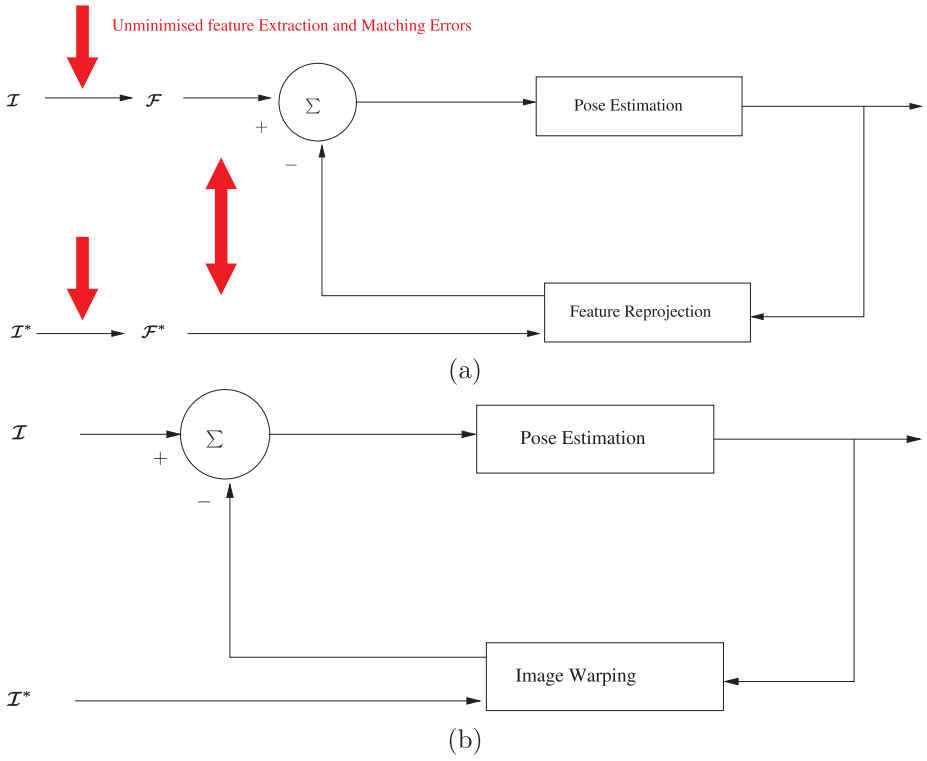


Fig. 1. Two non-linear iterative estimation loops. (a) A feature-based approach where it can seen that there unminimized errors made by the extraction of stereo features $\boldsymbol{\mathcal{F}}$ from both the current stereo images $\boldsymbol{\mathcal{I}}$ and the reference stereo images $\boldsymbol{\mathcal{I}}^*$. Furthermore, there are unminimized errors from the *temporal matching* between the reference stereo features and the current features. (b) A direct minimization approach that minimizes directly the pixel intensities between a warped image and the reference image. Here the errors shown in (a) are minimized. Note that in this paper, only the temporal correspondence errors are minimized and not the spatial errors. This is left as a perspective of the approach.

Stereo techniques provide accurate 3D information at little computational cost (1D search and matching along epipolar lines) and subsequently avoid the problems of monocular algorithms (i.e. scale factor, initialization, observability, etc.) by using prior knowledge about the extrinsic camera parameters and applying multi-view constraints. Indeed a multitude of work exists on multiview geometry (see Hartley and Zisserman (2001) and references therein). State-of-the-art stereo techniques (Nister et al. 2006; Howard 2008) are, however, only feature based and to the best of the authors' knowledge no work has been done on deriving an efficient direct tracker as in Benhimane and Malis (2004) and Silveira et al. (2008) using stereo warping and novel view synthesis as in Avidan and Shashua (2001).

Feature-based methods (Chiuso et al. 2002; Davison and Murray 2002; Nistér et al. 2004; Mouragnon et al. 2006; Konolige and Agrawal 2008) all rely on an intermediary estimation process based on detection thresholds. This feature extraction process is often badly conditioned, noisy and not robust therefore relying on higher-level robust estimation techniques. Since the global estimation loop is never closed on the image measurements (intensities) these multi-step techniques systematically propagate feature extraction and matching errors and accumulate drift (refer to Figure 1(a)). It is important to note that stereo-feature-based techniques require spatial matching across the stereo pair *and* temporal matching between stereo pairs. To eliminate drift these approaches resort to techniques such as local bundle adjustment or simultaneous localization and mapping (SLAM). Since feature-based techniques perform matching across the entire image, however, they have the advantage of being able to handle large inter-frame movements. Unfortunately, if there is little overlap between images, feature-based approaches are nonetheless prone to miss-matching and are unstable since there are fewer features to be matched and the estimator can become ill-conditioned.

Appearance, optical flow or direct techniques (Lucas and Kanade 1981; Irani and Anandan 2000), on the other hand, are image-based and minimize an error directly based on the image measurements (refer to Figure 1(b)). These approaches

have the advantage of being precise and perform tracking and pixel correspondence/matching simultaneously. Unfortunately, techniques published so far have only considered tracking between two monocular images (including spatial stereo matching) and often make heavy assumptions about the nature of the structure within the scene or the camera model. For example Hager and Belhumeur (1998) assumed an affine motion model and Baker and Matthews (2001) and Benhimane and Malis (2004) assumed planar homography models. In this way the perspective effects or the effects of non-planar 3D objects are not considered and tracking fails easily under large movements. Furthermore, these techniques all require the definition of a region of interest within the image and are limited to local convergence around that region. More specifically they require a *sufficient overlap* as opposed to feature-based techniques which can perform matching globally within the image.

Another very important issue is the registration problem. *Purely geometric* or *numerical and iterative* approaches may be considered. *Linear approaches* use a least-squares method to estimate the pose and are considered to be more suitable for initialization procedures. *Full-scale non-linear optimization techniques* (e.g. Hager and Belhumeur (1998), Baker and Matthews (2001), and Comport et al. (2006b)) consist of minimizing an objective function using numerical iterative algorithms such as Newton–Raphson or Levenberg–Marquardt. The main advantage of these approaches are their computational efficiency and their accuracy, however, they may be subject to local minima and, worse, divergence.

The technique proposed in this paper is a 3D visual odometry technique that minimizes a direct intensity error between consecutive stereo pairs. This approach lies at the intersection between direct image-based and model-based techniques. The image-based model is obtained by performing dense stereo matching either online when in unknown environments, or offline when a training set is available. The 3D model then comprises both photometric stereo image information along with a disparity map. The global image-based model can then be considered as a set of key reference image-pairs that are used to perform localization locally around those reference positions. Six-degree-of-freedom pose estimation is achieved by defining a quadrifocal warping function which closes a non-linear iterative estimation loop *directly* with the images. This approach handles arbitrary 3D structure and improves the convergence domain with respect to region-based methods since the entire image is used and the probability of a sufficient inter-frame overlap is therefore much higher. Whilst this approach improves the convergence domain, the accuracy of direct techniques (Irani and Anandan 2000) is retained since no feature extraction is performed. In terms of minimization, in this paper an efficient second-order approach (Benhimane and Malis 2004; Malis 2004) is employed which improves efficiency and also helps to avoid local minima.

As will be shown, the proposed technique is able to accurately handle large-scale scenes efficiently whilst avoiding error-prone feature extraction and inter-frame matching. This leads to impressive results in real scenes with occlusions, large inter-frame displacements, and very little drift. This paper is an extended and more detailed version of the technique presented in Comport et al. (2007). In Section 2 an overview of the objective function is given. In Section 3.2 the stereo warping function is detailed. Section 4 outlines the robust second-order minimization technique and the results are presented in Section 6.

2. Trajectory Estimation

A framework is described for estimating the trajectory of a stereo-camera rig along a sequence from a designated set of pixels within the image pair. The tracking problem will essentially be considered as a pose estimation problem which will be related directly to the gray-level brightness measurements within the stereo pair via a non-linear model which accounts for the dense 3D geometric configuration of the scene.

Since the ultimate objective is to control a robot within Euclidean space, a calibrated camera pair is considered. Consider a stereo camera pair with two brightness functions $\mathbf{I}(\mathbf{p}, t)$ and $\mathbf{I}'(\mathbf{p}', t)$ for the left and right cameras, respectively, where $\mathbf{p} = (u, v, 1)$ and $\mathbf{p}' = (u', v', 1)$ are homogeneous vectors containing the pixel locations within the two images acquired at time t. It is convenient to consider the set of image measurements in vector form such that $\boldsymbol{\mathcal{I}} = (\mathbf{I}, \mathbf{I}')^\top \in \mathbb{R}^{2n}$ is a vector of intensities of the left image stacked on top of the right, with n the number of pixels in one image.

Here $\boldsymbol{\mathcal{I}}$ will be called the *current* view pair and $\boldsymbol{\mathcal{I}}^*$ as the *reference* view pair. A superscript $*$ will be used throughout to designate the reference view variables. Similarly, with abuse of notation, $\boldsymbol{\mathcal{P}}^* = (\mathbf{p}^*, \mathbf{p}^{*\prime}) \in \mathbb{R}^4$, is a stereo image correspondence from the reference template pair. Any set of corresponding pixels from the reference image pair are considered as a reference template, denoted by $\mathcal{R}^* = \{\{\mathbf{p}^*_1, \mathbf{p}^{*\prime}_1\}, \{\mathbf{p}^*_2, \mathbf{p}^{*\prime}_2\}, \ldots, \{\mathbf{p}^*_n, \mathbf{p}^{*\prime}_n\}\}$ where n is the number of corresponding point pairs in the template.

The motion of the camera pair or objects within the scene induces a deformation of the reference template. The 3D geometric deformation of a stereo rig can be fully defined by a motion model $w(\boldsymbol{\mathcal{P}}^*, \mathbf{T}', \mathbf{K}, \mathbf{K}'; \overline{\mathbf{T}}(t))$. The motion model w considered in this paper is the quadrifocal warping function which will be detailed further in Section 3. Here $\mathbf{K}$ and $\mathbf{K}'$ contain the intrinsic calibration parameters for the left and right cameras, respectively. We use $\mathbf{T}' = (\mathbf{R}', \mathbf{t}') \in \mathrm{SE}(3)$ to denote the homogeneous matrix of the extrinsic camera pose of the right camera with respect to the left and $\overline{\mathbf{T}} = (\bar{\mathbf{R}}, \bar{\mathbf{t}}) \in \mathrm{SE}(3)$ is the current pose of the stereo rig relative to the reference position. Throughout, $\mathbf{R}$ is a rotation matrix and $\mathbf{t}$ the translation vector. Since both the intrinsic and extrinsic calibration parameters do not vary with time they will be assumed implicit.

It follows that the reference image is obtained by warping the current image as

$$\boldsymbol{\mathcal{I}}^*(\boldsymbol{\mathcal{P}}^*) = \boldsymbol{\mathcal{I}}\left(w\left(\boldsymbol{\mathcal{P}}^*; \overline{\mathbf{T}}\right)\right), \quad \text{for all } \boldsymbol{\mathcal{P}}^* \in \mathcal{R}^*, \tag{1}$$

where $\overline{\mathbf{T}}$ is the true pose. When the coordinates indexing the image do not correspond to an exact pixel location, bilinear interpolation is performed.

Suppose that at the current image an estimate of the pose $\widehat{\mathbf{T}}$ fully represents the pose of the stereo pair with respect to a pair of reference images. The tracking problem then becomes one of estimating the incremental pose $\mathbf{T}(\mathbf{x})$, where $\mathbf{x}$ is a minimal parametrization of the homogeneous pose matrix $\mathbf{T}$ and where it is supposed that there exists $\tilde{\mathbf{x}} : \widehat{\mathbf{T}}\mathbf{T}(\tilde{\mathbf{x}}) = \overline{\mathbf{T}}$. The estimate is updated by a homogeneous transformation $\widehat{\mathbf{T}} \leftarrow \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})$. It can be noted here that the increment is now parametrized to the right of $\widehat{\mathbf{T}}$ as opposed to the approach in Comport et al. (2007) where it is parametrized to the left ($\mathbf{T}_L(\mathbf{x})$). Both cases are related by

$$\mathbf{T}(\mathbf{x}) = \widehat{\mathbf{T}}^{-1}\mathbf{T}_L(\mathbf{x})\widehat{\mathbf{T}}. \tag{2}$$

This allows us to simplify the derivation of the Jacobian (described in Appendix A) so that a pre-calculation of the Jacobian can be made to improve the computational efficiency.

The unknown parameters $\mathbf{x} \in \mathbb{R}^6$ are defined by the Lie algebra $\mathfrak{se}$ as

$$\mathbf{x} = (\boldsymbol{\omega}\Delta t, \boldsymbol{v}\Delta t) \in \mathfrak{se}, \tag{3}$$

which is the integral of a constant velocity twist which produces a pose $\mathbf{T}$, where $\boldsymbol{v}$ and $\boldsymbol{\omega}$ are the linear and angular velocities, respectively. The pose and the twist are related via the exponential map as

$$\mathbf{T}(\mathbf{x}) = \exp\left(\begin{bmatrix} [\boldsymbol{\omega}]_\times & \boldsymbol{v} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right), \tag{4}$$

where $[\cdot]_\times$ represents the skew symmetric matrix operator.

Thus, the pose and the trajectory of the camera pair can be estimated by minimizing a non-linear least-squares cost function:

$$C(\mathbf{x}) = \sum_{\boldsymbol{\mathcal{P}}^* \in \mathcal{R}^*} \left(\boldsymbol{\mathcal{I}}\left(w(\boldsymbol{\mathcal{P}}^*; \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}))\right) - \boldsymbol{\mathcal{I}}^*\left(\boldsymbol{\mathcal{P}}^*\right)\right)^2. \tag{5}$$

This function is minimized using the robust, efficient and precise second-order minimization procedure detailed in Section 4.

3. Novel View Synthesis and Warping

The geometric configuration of a stereo pair, that is undergoing movement within a rigid scene, is based on the paradigm that four views of a scene satisfy quadrifocal constraints. Thus, given a reference stereo view with correspondences between pixels and the quadrifocal tensor, a third view and fourth view can be generated by means of a warping function. This warping function subsequently provides the required relationship between two views of the scene and an adjacent pair of views in a sequence of images.

The approach presented here is formalized using the quadrifocal tensor since it encapsulates all of the geometric relations between four views that are independent of scene structure and provides a clear insight into the geometric properties (homography transfer between two views via a line in the third, point–line relations, etc.). This allows us to clearly define a choice of quadrilinear constraints from the full set of possible constraints and to identify any degenerate configurations. In this way a clear link is made with projective multi-view geometry and extensions to our approach are therefore facilitated (for example, the extension to a fully projective approach). Furthermore, the analytical development of the image warping function ensures that the measurement uncertainties are consistently handled in the optimization procedure developed in Section 4.

3.1. Quadrifocal Geometry

A point $\mathbf{X} \in \mathbb{R}^3$ in Euclidean space projects onto the 3D camera plane by a 3×4 projection matrix $\mathbf{M} = \mathbf{K}[\mathbf{R}|\mathbf{t}] \in \mathbb{P}^3$ where the image point is given by $\mathbf{p} = \mathbf{M}\mathbf{X}$ so that $\mathbf{p} = (u, v, 1)^\top$ is the homogeneous pixel vector (see Figure 2). It is assumed that the stereo rig is calibrated with intrinsic camera parameters $\mathbf{K}$ and $\mathbf{K}'$ for the left and right cameras, respectively, and extrinsic parameters $\mathbf{T}'$ denoting the pose from the left to right camera.

Much work has been carried out on multi-view geometry and of particular interest here are the quadrilinear relations (Faugeras and Mourrain 1995; Hartley 1995; Triggs 1995; Heyden and Astrom 1997; Shashua and Wolf 2000) between two pairs of stereo images at two consecutive time instants. The compact tensor notation of multi-focal geometry will be used here with a covariant–contravariant summation convention. Contravariant point vectors $\mathbf{p}^i$ are denoted with a superscript and their covariant counterpart representing lines $\mathbf{l}_j \in \mathbb{P}^2$, are denoted with a subscript. A contraction or summation over two tensors occurs when there are repeated indices in both contravariant and covariant variables (i.e. $\mathbf{p}^i\mathbf{l}_i = \sum_{j=1}^n \mathbf{p}^j\mathbf{l}_j$). An outer-product of two first-order tensors (vectors), $\mathbf{a}_i\mathbf{b}^j$ is a second-order tensor (matrix) $\mathbf{c}_i^j$ which is equivalent to $\mathbf{C} = \mathbf{b}\mathbf{a}^\top$ in matrix notation.

Consider a point in correspondence across four views, $\mathbf{p} \leftrightarrow \mathbf{p}' \leftrightarrow \mathbf{p}'' \leftrightarrow \mathbf{p}'''$, with the camera matrices $\mathbf{M}, \mathbf{M}', \mathbf{M}'', \mathbf{M}'''$. A common method for deriving the quadrifocal tensor is then to define the linear system as

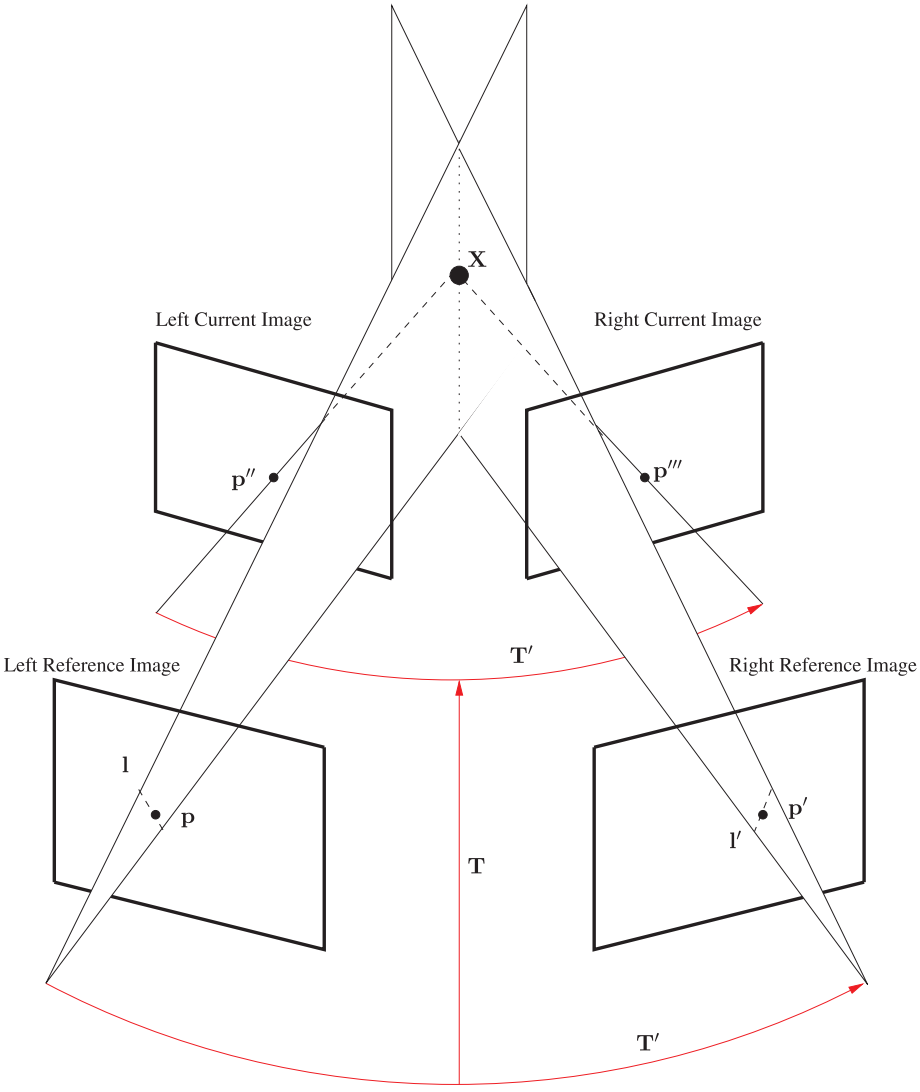


Fig. 2. The quadrifocal geometry of a stereo pair at two subsequent time instants. Two points $\mathbf{p}$ and $\mathbf{p}'$ are initialized only once at the beginning of the tracking process to be in correspondence. The central pose $\mathbf{T}$ is estimated via a non-linear warping function which warps all points in the reference stereo pair to the current image points $\mathbf{p}''$ and $\mathbf{p}'''$. The quadrifocal warping function is defined by choosing two lines $\mathbf{l}$ and $\mathbf{l}'$ passing through corresponding points in the first image. The extrinsic parameters $\mathbf{T}'$ are assumed known *a priori*.

$$\begin{bmatrix} \mathbf{M} & \mathbf{p} & & & \\ \mathbf{M}' & & \mathbf{p}' & & \\ \mathbf{M}'' & & & \mathbf{p}'' & \\ \mathbf{M}''' & & & & \mathbf{p}''' \end{bmatrix} \begin{pmatrix} \mathbf{X} \\ -k \\ -k' \\ -k'' \\ -k''' \end{pmatrix} = \mathbf{0}, \qquad (6)$$

where k, k', k'', k''' are unknown scale factors.

The matrix forming the left-hand part of expression (6) is of dimension 12×8. Owing to the existence of a solution and the fact that the right term is not zero, any 8×8 minor has a zero determinant. This fact defines the quadrilinear relations between the points in the four views. From the choice of the eight rows from the various camera matrices two different cases may be considered:

1. The case where eight rows are selected such that only one row from a camera matrix is included in the deter-

minant leads to a trilinear or bilinear relationship. This case is not considered here.

2. The case where two rows from each camera matrix is included leads to a quadrifocal relationship, which is considered here.

The quadrifocal tensor that transfers a point in the left reference view, that is, in correspondence with a point in the right reference view, to points in the current left and right views is then written as

$$\mathbf{p}''^{r}\mathbf{p}'''^{s} = \mathbf{p}^{i}\mathbf{p}'^{j}\epsilon_{ipw}\epsilon_{jqx}\mathcal{Q}^{pqrs}, \tag{7}$$

where ϵ is the permutation tensor with properties that can be used to represent the vector cross product, the skew symmetric matrix and the determinant. It is provided here for completeness:

$$\epsilon_{ijk} = \begin{cases} 0, & \text{unless } i, j, k \text{ are distinct,} \\ +1, & \text{if } i, j, k \text{ are an even permutation of } 1, 2, 3, \\ -1, & \text{if } i, j, k \text{ are an odd permutation of } 1, 2, 3. \end{cases} \tag{8}$$

The quadrifocal tensor is a fourth-order tensor represented by a homogeneous $3\times3\times3\times3$ array of elements. It is obtained by expanding the determinant of the matrix in Equation (6) in terms of the points $\mathbf{p}, \ldots, \mathbf{p}'''$ as

$$\mathbf{Q}^{pqrs} = \det \begin{bmatrix} \mathbf{m}^{p} \\ \mathbf{m}'^{q} \\ \mathbf{m}''^{r} \\ \mathbf{m}'''^{s} \end{bmatrix}, \tag{9}$$

where the contravariant indices p, q, r, s index the rows of the camera matrices $\mathbf{M}, \ldots, \mathbf{M}'''$.

In order to transform the points in the reference images to points in the current images it is possible to either use a single quadrifocal tensor or to define two quadrifocal tensors, one for each reference image. This choice depends on the stereo matching precision (sub-pixel or one-to-one). First define the left and right quadrifocal tensors as $\mathbf{Q}_L$ and $\mathbf{Q}_R$. These two cases are then as follows:

1. *Single quadrifocal tensor*. This case corresponds to a one-to-one matching between the reference images so that $\mathbf{Q}_L = \mathbf{Q}_R$. The disadvantage here is that the correspondences between left and right reference images are only approximate and to not have sub-pixel matching accuracy. Furthermore, most one-to-one dense correspondence algorithms do not give the same results in each direction. The advantage is that it is computationally twice as fast as case 2.
2. *Two quadrifocal tensors*. This case corresponds to performing sub-pixel matching for both left and right images, however, the disadvantage is that there are twice as many constraints to be applied.

These two cases are also due to the wish to perform *direct* minimization with the image measurements of both cameras. It can be noted, however, that if the estimation loop is not closed *directly* with both images, then it is possible to perform sub-pixel matching with the left image as the reference (for instance) and to approximate the right image by interpolating the corresponding pixels so as to maintain a one-to-one mapping. In this case the computation remains efficient and sub-pixel matching is performed, however, the estimation is biased towards one image (i.e. the error being minimized is only direct with respect to the left image). This approximation can subsequently lead to inaccurate results whether it be inaccurate correspondences or a bias with respect to errors made in the intrinsic parameters of a "dominant" camera.

In this paper case 2 will be developed further and the following paragraphs will show how this dual relationship can be simplified into a single quadrifocal tensor whilst maintaining symmetry. Shashua and Wolf (2000) gave a detailed method for decomposing the quadrifocal tensor as an epipole–homography pairing. Of interest here is the composition of the quadrifocal tensor from two trifocal and a bifocal relationship:

$$\delta_i\mu_j\mathbf{Q}^{ijkl} = [\delta_i\mu_j\mathcal{T}_l^{ij}]_x\mathbf{F}_{12}[\delta_i\mu_j\mathcal{T}_k^{ij}]_x, \tag{10}$$

where x is an index from the set i, j, k, l and δ_i and μ_j range over the standard bases $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$. Under this definition every $3\times3\times3$ slice of $\mathbf{Q}^{ijkl}$ corresponds to a homography tensor Shashua and Wolf (2000) between the remaining views not represented by x.

Following from the previous discussion, each of the two sub-pixel quadrifocal tensors provides a relationship between corresponding points in the four images that can be decomposed into two trifocal tensors and a fundamental matrix. This decomposition will be used here to reduce the number of quadrilinear relations to a single set, whilst maintaining the one-to-many correspondences for both the left and right reference images. With the given decomposition (and before simplification), this makes a total of two fundamental matrices and four trifocal tensors:

1. Two fundamental matrices between the left and right reference images (one in each direction): since the pose between the reference and current viewpoints is initially unknown then the transfer of points between the left and right reference views is a bilinear relation that depends on the matched points between the two images. It is assumed here that the matching is performed *a priori* by a dense correspondence technique that exploits this bilinear relation via a 1D search along epipolar lines.

2. Two trifocal tensors relating each reference image to its current image via a corresponding reference image: this is the most important case as it depends on the unknown pose and remains robust to camera modeling errors.
3. Two trifocal tensors relating each reference image to the opposite current image: this case is interesting since it depends on the unknown pose, however, this case is more sensitive to camera calibration or modeling errors. This could be a good candidate if one wished to estimate the calibration parameters.

In order to reduce both quadrifocal tensors into a single relation, the left–right fundamental matrix is retained along with the two trifocal tensors given by case 2 above. In this way the quadrilinear constraints are maintained while reducing the computational complexity of the optimization. Thus, from Equation (10), $\mathcal{T}_l^{ij}$ is chosen as the trifocal tensor between views (4,1,2), $\mathcal{T}_k^{ij}$ is chosen as the trifocal tensor between views (3,1,2) and $\mathbf{F}_{12}$ the fundamental matrix between views (1,2).

The geometry between two stereo pairs is therefore defined in a manner that is simple for subsequent developments using the canonical coordinates of two triplets of images. First of all, consider the triplet consisting of the left reference camera, the right reference camera and the left current camera. The left reference camera matrix is chosen as the origin so that $\mathbf{M} = \mathbf{K}[\mathbf{I}|\mathbf{0}]$. The reference projection matrix for the right camera (the extrinsic camera pose) and the *current* projection matrices for the left camera are then $\mathbf{M}' = \mathbf{K}'[\mathbf{R}'|\mathbf{t}']$ and $\mathbf{M}'' = \mathbf{K}[\mathbf{R}''|\mathbf{t}'']$.

The second triplet is defined in a similar manner such that the right reference camera is chosen as the origin and the left reference camera and right current camera matrices are defined with respect to this origin.

In order to construct the quadrifocal relation it is necessary to combine these two triplets of images using the left–right bilinear relation. This is done symmetrically by defining the arbitrary *world origin* as the geodesic center between the two reference cameras. To do this the extrinsic parameters are separated into two distinct poses with respect to the center as

$$\mathbf{T}^c = \exp^{(\log(\mathbf{T}'))/2} \quad \text{and} \quad \mathbf{T}'^c = \mathbf{T}^c\mathbf{T}'^{-1}, \tag{11}$$

where exp and log are the matrix exponential and logarithm.

The pose from the left reference camera to the current one is therefore composed of a central pose as

$$\mathbf{T}'' = \mathbf{T}^{c\,-1}\widetilde{\mathbf{T}}\mathbf{T}^c, \tag{12}$$

where $\widetilde{\mathbf{T}}$ is the unknown pose to be estimated. In Section 4 we show how this pose may be estimated iteratively.

3.2. Quadrifocal Warping

The quadrifocal warping function $w(\boldsymbol{\mathcal{P}}^*; \overline{\mathbf{T}})$ from Equation (5) can now be considered to be composed of a trifocal tensor for each of the left and right images, that depend on the unknown minimal pose parameters, along with the constant extrinsic camera pose that provides the bilinear constraint of Equation (10). Since, the bilinear constraint corresponds to a constant change of reference frame, this will be performed during the estimation in Section 4. For the moment the focus will be on defining the warping of each left and right image via their corresponding trifocal tensors. In overview, the trifocal tensor is used to transfer (warp) corresponding points from two views to a third view. This tensor depends only on the relative motion of the cameras as well as the intrinsic and extrinsic camera parameters.

The trifocal tensor $\mathcal{T}$ is a third-order tensor represented by a homogeneous $3 \times 3 \times 3$ array of elements. The trifocal tensor can be determined from equation (9) by taking into account the case of one line of the last camera matrix. The calibrated case is given as:

$$\begin{aligned}\mathcal{T}_i^{jk} &= \mathbf{k}'^{j}_{m}\mathbf{r}'^{m}_{n}\mathbf{k}^{-1}{}^{n}_{i}\mathbf{k}''^{k}_{o}\mathbf{t}''^{o}(t) \\ &\quad - \mathbf{k}'^{j}_{p}\mathbf{t}'^{p}\mathbf{k}''^{k}_{q}\mathbf{r}''^{q}_{r}(t)\mathbf{k}^{-1}{}^{r}_{i},\end{aligned} \tag{13}$$

where $(\mathbf{r}', \mathbf{t}')$ and $(\mathbf{r}'', \mathbf{t}'')$ are the tensor forms of the rotation matrix and translation vector for the second and third camera matrices, respectively. Here $\mathbf{k}$ and $\mathbf{k}'$ are the intrinsic calibration components of the left and right camera matrices, respectively. Note that $\mathbf{k}'' = \mathbf{k}$ or $\mathbf{k}'' = \mathbf{k}'$ and that the roles of left and right cameras are shifted depending on whether one is warping to the left or right camera at the next time instant.

Given any line $\mathbf{l}$ coincident with $\mathbf{p}$ or any line $\mathbf{l}'$ coincident with $\mathbf{p}'$, then the trifocal tensor contracts so as to become a homography $\mathbf{h}$ which maps points from one reference image to the current image, i.e. a line defined in one of the reference views defines a plane which can be used to warp a point between the remaining reference image and the current image. Thus, the warping from the left reference image to the left current image via a plane in the right reference image is given by

$$\mathbf{p}''^{k} = \mathbf{p}^i\mathbf{l}'_j\mathcal{T}_i^{jk} = \mathbf{h}_i^k\mathbf{p}^i,$$

where $\mathbf{p}^i$ is a point in the left reference image, $\mathbf{l}_k$ is a line defined in the right reference image and $\mathbf{p}''^{k}$ is the warped point in the left current image. This equation is used similarly for warping a point in the right reference image to a point in the right current via a plane in the left reference image.

As opposed to transfer using the fundamental matrix, the tensor approach is free from singularities when the 3D point lies on the trifocal plane. The only degenerate situation that occurs is if a 3D point lies on the baseline joining the first and the second cameras since the rays through $\mathbf{p}$ and $\mathbf{p}'$ are co-linear.

It is, however, important to carefully choose the seemingly arbitrary lines passing through the points $\mathbf{p}$ and $\mathbf{p}'$. In particular, the trifocal tensor is not defined when the epipolar

line is chosen. The lines may, however, be chosen in several ways (Hartley and Zisserman 2001) including an optimal least-squares solution to the linear system of equations $\mathbf{p}^i(\mathbf{p}'^j\epsilon_{jpr})(\mathbf{p}''^k\epsilon_{kqs})\mathcal{T}_i^{pq} = \mathbf{0}_{rs}$, where ϵ is the tensor which transforms $\mathbf{p}$ into its skew-symmetric form $[\mathbf{p}]_\times$. This choice, however, immensely complicates the analytic derivation of the Jacobian given in Section 4. Other alternatives include choosing the line perpendicular to the epipolar line, or computing the result using several lines and choosing the best. In the case of a calibrated stereo rig the epipolar geometry is known so it is possible to directly choose a single line that is not degenerate and at the same time one that simplifies the derivation of the Jacobian. Subsequently, the reference line is chosen to be the diagonal line $\mathbf{l} = (-1, -1, u+v)$ coincident with the point (u, v).

The stereo warping operator is then given by

$$\begin{bmatrix} \mathbf{p}''^k \\ \mathbf{p}'''^n \end{bmatrix} = \begin{bmatrix} \mathbf{p}^i\mathbf{l}'_j\mathcal{T}_i^{jk} \\ \mathbf{p}'^l\mathbf{l}_m\mathcal{T}_l^{mn} \end{bmatrix}, \tag{14}$$

where the indexes of the two trifocal tensors indicate tensors transferring to the left and right cameras. The lines $\mathbf{l}'$ and $\mathbf{l}$ are chosen to be the diagonal line as outlined in the previous paragraph. If these trifocal tensors are contracted into constant $(\mathbf{p}, \mathbf{p}', \mathbf{l}, \mathbf{l}', \mathbf{r}', \mathbf{t}', \mathbf{k}, \mathbf{k}')$ and non-constant components $(\mathbf{r}'', \mathbf{t}'')$, the warping function is composed of projective 3D points (expanded with the current calibration parameters) and the unknown pose.

It is important for further developments to highlight that the warping operator $w(\boldsymbol{\mathcal{P}}^*; \overline{\mathbf{T}}) : \mathrm{SE}(3) \times \mathbb{R}^4 \longrightarrow \mathbb{R}^4$ is a *group action*. Indeed, the following operations hold:

1. The identity map:

$$w(\boldsymbol{\mathcal{P}}^*; \mathbf{I}) = \boldsymbol{\mathcal{P}}^*, \quad \text{for all } \boldsymbol{\mathcal{P}}^* \in \mathbb{R}^4, \tag{15}$$

2. The composition of an action corresponds to the action of a composition for all $\mathbf{T_1}, \mathbf{T_2} \in \mathrm{SE}(3)$:

$$w(w(\boldsymbol{\mathcal{P}}^*, \mathbf{T_1}), \mathbf{T_2}) = w(\boldsymbol{\mathcal{P}}^*, \mathbf{T_1}\mathbf{T_2}) \quad \text{for all } \boldsymbol{\mathcal{P}}^* \in \mathbb{R}^4. \tag{16}$$

4. Robust and Efficient Quadrilinear Tracking

The aim now is to minimize the difference in image intensities from the objective criterion defined previously (5) in an accurate and robust manner. If the L_2 norm is chosen the approach is well known as the sum-of-squared difference (SSD) tracking. If a robust objective function is considered then the objective function therefore becomes

$$O(\mathbf{x}) = \sum_{\boldsymbol{\mathcal{P}}^* \in \mathcal{R}^*} \rho\left(\boldsymbol{\mathcal{I}}\left(w(\boldsymbol{\mathcal{P}}^*; \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}))\right) - \boldsymbol{\mathcal{I}}^*(\boldsymbol{\mathcal{P}}^*)\right), \tag{17}$$

where $\rho(u)$ is a robust function (Huber 1981) that grows subquadratically and is monotonically non-decreasing with increasing $|u|$ (see Appendix C).

Since this is a non-linear function of the unknown pose parameters an iterative minimization procedure is employed. The robust objective function is minimized by $[\nabla O(\mathbf{x})]_{\mathbf{x}=\tilde{\mathbf{x}}} = 0$, where ∇ is the gradient operator with respect to the unknown parameters (3) and there exists a stationary point $\mathbf{x} = \tilde{\mathbf{x}}$ which is the global minimum of the cost function.

Pseudo-second-order methods such as Levenberg–Marquardt are generally employed to minimize iteratively such an objective function since computing the Hessian to obtain full quadratic convergence is computationally expensive. However, since both the reference image and current image are available, along with the fact that the warping operator has group properties, it is possible to use the efficient second-order approximation (ESM) proposed by Malis (2004) and Benhimane and Malis (2004). This technique allows us to avoid the computation of the Hessian.

For completeness the essential steps are summarized here. Consider the general least-squares minimization problem:

$$F(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{n}(f_i(\mathbf{x}))^2 = \frac{1}{2}\|\mathbf{f}(\mathbf{x})\|^2. \tag{18}$$

In order to minimize this non-linear function an iterative gradient descent is performed in order to search for $\nabla F(\tilde{\mathbf{x}}) = \mathbf{0}$, where the definition of ∇ is taken to be the multi-variate, multi-function Jacobian gradient operator which takes the derivative of $F(\mathbf{x})$ with respect to $\mathbf{x}$. To derive the ESM, the first step is a second-order Taylor series expansion of $\mathbf{f}$ about $\mathbf{x} = \mathbf{a}$ as

$$\mathbf{f}_i(\mathbf{x}) = \mathbf{f}_i(\mathbf{a}) + \nabla_i^j\mathbf{f}(\mathbf{a})(\mathbf{x}_j - \mathbf{a}_j) \tag{19}$$

$$+ \frac{1}{2}\nabla_i^{jk}\mathbf{f}(\mathbf{a})(\mathbf{x}_j - \mathbf{a}_j)(\mathbf{x}_k - \mathbf{a}_k) + \mathcal{R}_1(\|\mathbf{x}\|^3), \tag{20}$$

where the Jacobian is $\mathbf{J}_i^j(\mathbf{x}) = \nabla_i^j\mathbf{f}(\mathbf{x})$ is an second-order tensor of dimension $n \times 6$, the Hessian tensor is a third-order tensor $\mathbf{H}_i^{jk}(\mathbf{x}) = \nabla_i^{jk}\mathbf{f}(\mathbf{x})$ of dimension $n \times 6 \times 6$ and $\mathcal{R}_1(\|\mathbf{x}\|^3)$ is the third-order remainder. The Jacobian $\mathbf{J}$ can also be approximated via a Taylor series expansion as

$$\mathbf{J}_i^j(\mathbf{x}) = \mathbf{J}_i^j(\mathbf{a}) + \mathbf{H}_i^{jk}(\mathbf{a})(\mathbf{x}_k - \mathbf{a}_k) + \mathcal{R}_2(\|\mathbf{x}\|^2). \tag{21}$$

Substituting for $\mathbf{H}_i^{jk}(\mathbf{a})(\mathbf{x}_k - \mathbf{a}_k)$ from Equation (21) into Equation (20) and evaluating at $\mathbf{a} = \mathbf{0}$ gives

$$\mathbf{f}_i(\mathbf{x}) = \mathbf{f}_i(\mathbf{0}) + \frac{1}{2}\left(\mathbf{J}_i^j(\mathbf{0}) + \mathbf{J}_i^j(\mathbf{x})\right)\mathbf{x}_j + \mathcal{R}_3(\|\mathbf{x}^3\|). \tag{22}$$

It can be seen here that if the third-order terms are ignored, the second-order approximation depends on both the Jacobian evaluated at the current position $\mathbf{J}(\mathbf{0})$ and the Jacobian evaluated at the solution $\mathbf{J}(\mathbf{x})$. Since $\mathbf{x}$ is the unknown solution to

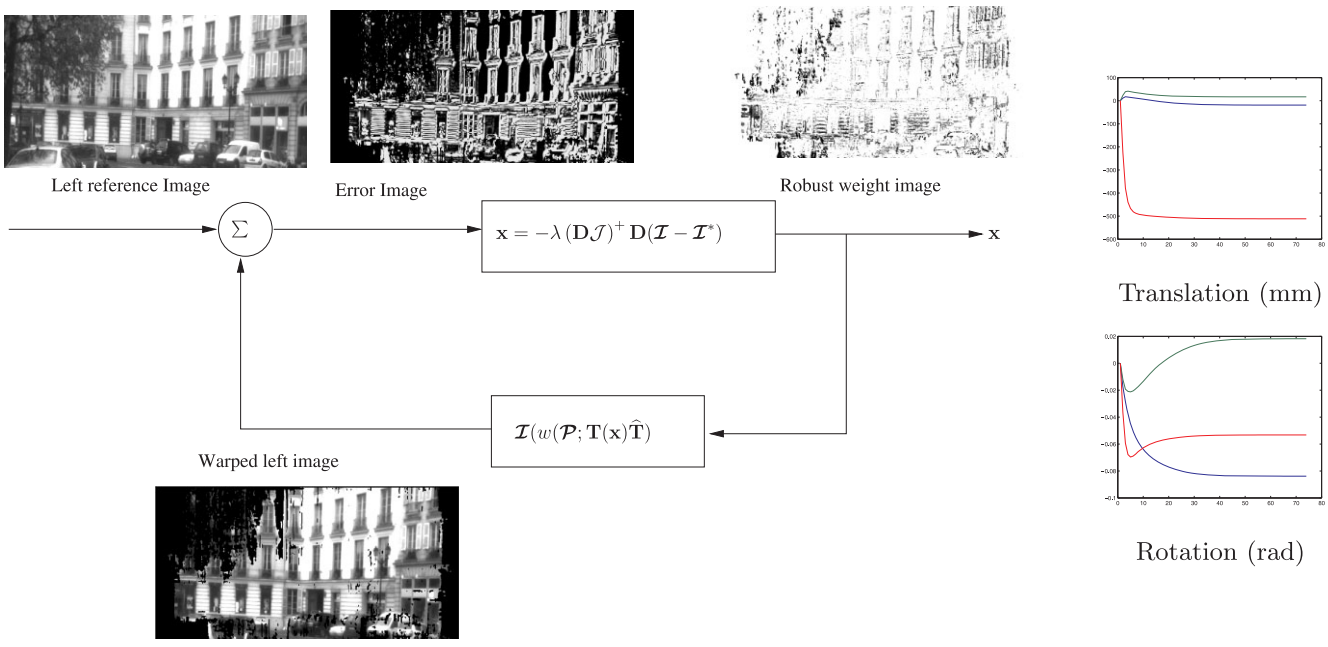


Fig. 3. The iterative estimation process as given by Equation (25). The reference image is given as input. An error image is then obtained with the warped current image. The robust second-order term is inversed to obtain an incremental pose (parametrized as an element of the Lie algebra here). The current image is then re-warped and the process is repeated until convergence. Only the left images and corresponding data are given here, however, the same is available for the right image. The smooth convergence of the translation and rotation is shown here for a large movement without prediction and at full resolution.

the system of equations, it would seem impossible to determine this term. However, it will be shown that in the present tracking case, it is possible to substitute the current image in Equation (17) for the reference image (image at the solution) in order to obtain an equivalent term without knowing $\mathbf{x}$.

In the case of the stereo image function, the second-order expansion is then given as

$$\mathcal{I}(\tilde{\mathbf{x}}) \approx \mathcal{I}(\mathbf{0}) + \frac{\mathbf{J}(\mathbf{0}) + \mathbf{J}(\tilde{\mathbf{x}})}{2}\tilde{\mathbf{x}}, \tag{23}$$

where $\mathbf{J}(\mathbf{0})$ is the current image Jacobian and $\mathbf{J}(\tilde{\mathbf{x}})$ is reference image Jacobian.

The current Jacobian and the reference Jacobians can be decomposed as the product of four Jacobians. Their detailed derivation can be found in Appendix A. In summary of the discussion found in the appendix, the second-order approximation of Equation (23) gives

$$\mathcal{J}(\tilde{\mathbf{x}}) = \frac{(\mathbf{J}_{\mathcal{I}} + \mathbf{J}_{\mathcal{I}^{*}})}{2}\mathbf{J}_{\mathbf{w}}\mathbf{J}_{\mathbf{T}}\mathbf{J}_{\mathcal{V}}, \tag{24}$$

where only $\mathbf{J}_{\mathcal{I}}$ varies with time and needs to be computed at each iteration.

The objective function is minimized by iteratively solving Equation (17) by using Equations (24) and (14) to obtain

$$\tilde{\mathbf{x}} = -\lambda\,(\mathbf{D}\mathcal{J})^{+}\,\mathbf{D}(\mathcal{I} - \mathcal{I}^{*}), \tag{25}$$

where $(\mathbf{D}\mathcal{J})^{+}$ is the pseudo-inverse, $\mathbf{D}$ is a diagonal weighting matrix determined from a robust function (see Appendix C) and λ is the gain which ensures an exponential decrease of the error. Refer to Figure 3 for a summary of this estimation process.

5. Implementation

5.1. Multi-resolution Tracking

In order to improve the computational efficiency of the approach and to handle large displacements a multi-resolution reference image (Burt 1984; Odobez and Bouthemy 1995) was constructed and used for tracking (refer to Figure 4). As is commonly done in this type of approach, the tracking begins at the highest levels (the lowest resolution) and performs tracking at this level until convergence. The ith resolution image is obtained by simply warping the original images as

$$\mathcal{I}_i = \mathcal{I}(w_H(\mathcal{P}, \mathbf{H}_i)), \tag{26}$$

where w_H is a stereo homographic warping function that warps the points $\mathcal{P}$ using the 3×3 homography $\mathbf{H}_i$ as

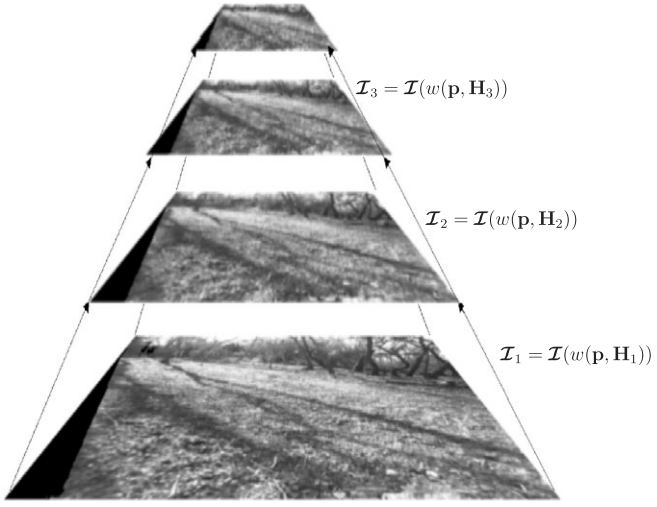


Fig. 4. A multi-resolution pyramid used to improve the computational performance of the tracking and avoid local minima. The homographic transform between scales is given in Figure 26.

$$\mathbf{p}_i = \begin{bmatrix} scaleu_i & 0 & 0 \\ 0 & scalev_i & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{p}. \tag{27}$$

Note that this is equivalent to changing the calibration parameters in the stereo warping function of Equation (14).

It should be noted here that the Nyquist–Shannon image re-sampling theorem is a widely studied problem in image processing (Howard 1983; Burt 1984) and computer graphics (Heckbert 1989) and it is well known that aliasing effects can occur if this is not done correctly. In general Gaussian smoothing is performed to eliminate aliasing but here we use simple bilinear interpolation which acts as a local box filter that provides a reasonable trade-off between computationally efficient filtering and aliasing.

Once convergence is obtained the current pose is used to initialize the next level in the pyramid and tracking is once again performed until convergence. This is repeated until the highest resolution of the pyramid is reached. In this way the larger displacements are minimized at lower cost on smaller images. Furthermore, the smaller images smooth out much of the detail required for fine adjustment and provide more global information that helps target the larger movements initially. This also has the effect of avoiding certain local minima.

5.2. Reference Image Pairs

As the camera pair moves through the scene the reference image may no longer be visible or the warped resolution becomes so poor that it is necessary to interpolate many pixels. In both cases this leads to mis-tracking. Therefore, in order to perform large-scale tracking it is necessary to continually update the reference image pair $\boldsymbol{\mathcal{I}}^*$. An update is detected by monitoring the error norm along with a robust estimate of the scale of the error distribution (i.e. the median absolute deviation). As soon as they become too large another set of dense correspondences between the stereo pair is made so as to reinitialize the tracking. As long as the same reference image is used then the minimization cut-off thresholds can be tuned for speed since the next estimation will recover any remaining error, however, if the reference image is changed the previous estimate is minimized with smaller cut-off thresholds so as to minimize any drift that may be left over.

5.3. Dense Correspondences

As mentioned, the reference image pairs need to be initialized with dense correspondences. The correspondence problem has been heavily studied in the computer vision literature and many different approaches are possible (Scharstein et al. 2001). When the cameras are calibrated the correspondence problem reduces to a 1D search along epipolar lines. This can either be performed off-line in a learning phase or on-line at each new acquisition depending on computational requirements (real-time approaches are feasible (van der Mark and Gavrila 2006)). In this paper the approaches given in Scharstein et al. (2001) and Ogale and Aloimonos (2005) along with a custom real-time GPU implementation for basic SSD correlation were used and tested. Nevertheless, any other type of dense correspondence algorithm could be used. The method given in Ogale and Aloimonos (2005) was initially used since it is particularly suited to urban canyon environments since the notions of horizontal and vertical slant are used to approximate first-order piecewise continuity. In this way the geometric projection of slanted surfaces from N pixels on one epipolar line to M pixels on another is not necessarily one-to-one but can be many-to-one or one-to-many. See Figure 5(c) for correspondence results of a typical image pair. In this case the disparity search region was fixed in a range of -20 to -180 pixels along the epipolar lines. In Figure 5(d) the visual quality of the results can be inspected. In this case the points in the right reference image are warped to the left by interpolation and it can be seen that the results is quite similar to the original image in Figure 5(b) apart from the fact that there are occluded regions in black where the dense matching algorithm did not succeed.

5.4. Robust Estimation

A robust M-estimation technique (as detailed in Appendix C and Comport et al. (2006a)) was used to reject outliers not corresponding to the definition of the objective function. The use

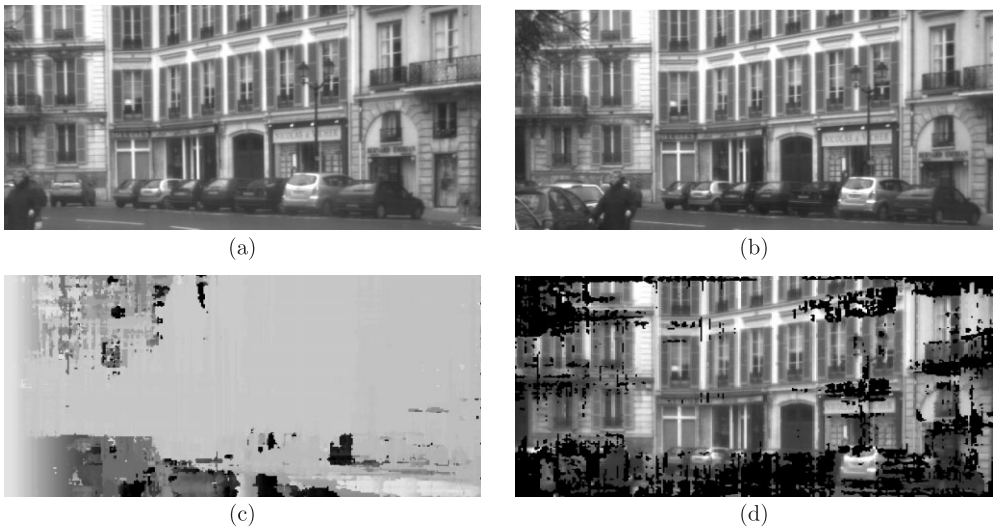


Fig. 5. Dense correspondence of an urban canyon with correspondence occlusion in black: (a),(b) right and left original images, respectively; (c) the disparity image from left to right; and (d) right image warped to the left image using the disparity values with occluded disparities in black.

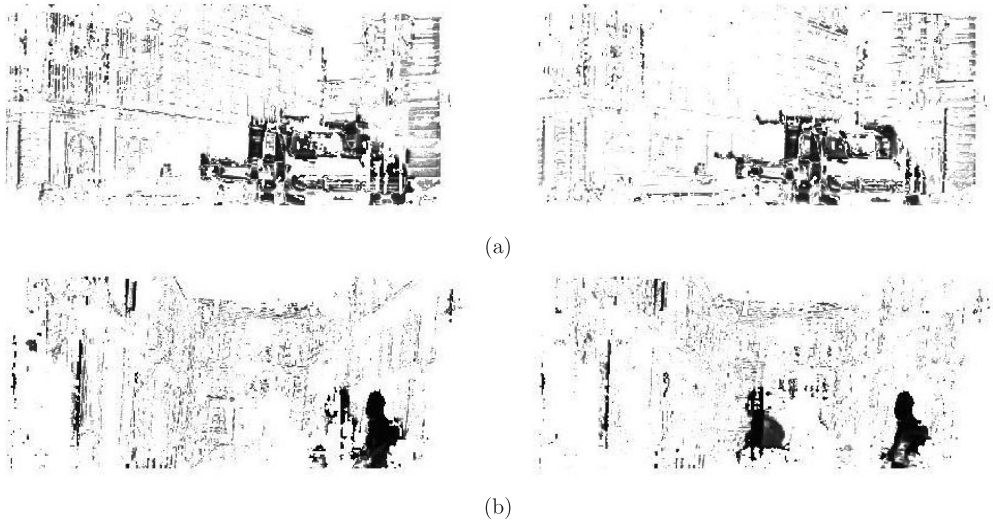


Fig. 6. Robust outlier rejection: two images showing the outlier rejection weights. The darker the points, the less influence they have on the estimation process. In (a) it can be seen that a moving truck has been rejected. In (b) a moving pedestrian has been rejected. It can also be noted that other outliers are detected in the image. These points generally correspond to matching error, noise in the image or the self-occlusion of the corners of the buildings.

of robust techniques is very interesting in the case of a highly redundant set of measurement as is the case of a set of dense correspondences. The outliers generally correspond to occlusions, illumination changes, matching error, noise in the image or the self-occlusion of the corners of the buildings.

In Figure 6(a) a moving truck has been rejected as an outlier whilst a stationary truck in the background was used to estimate the pose. In this way it can be seen that the proposed algorithm exploits all of the rigid information in the image so as to estimate the pose. In Figure 6(b) a moving pedestrian has been rejected and it can be seen that both the pedestrian projected from the reference image as well as the current position of the pedestrian have been rejected. This type of information has proven to be useful in an ap-

plication for initializing and tracking the trajectory of moving obstacles.

5.5. Extended Kalman Filter

In the context of tracking the trajectory of a car, very large inter-frame movements are observed. In the sequences considered in the following results, typical inter-frame movement was 1–2 m per image with a car travelling between 50 and 70 km h^{-1}. Even though tracking succeeds without predictive filtering, in order to improve computational efficiency (significantly less iterations in the minimization) a predictive filter was used. In this paper the well-known extended Kalman filter described by Zhang and Faugeras (1992) was used to predict the pose, however, the final pose estimate shown in the results was not filtered (i.e. the motion model was given no confidence and the filter was only used to initialize the estimator).

6. Results

6.1. Simulation Results

In order to test the algorithm with a ground truth a synthetic video sequence was created by warping real images onto various 3D surfaces. In this way realistic images were created with a known ground truth about the trajectory of the camera along with knowledge of the true image correspondences and camera calibration parameters.

In Figure 7, a 3D sphere is considered. A patch is selected on the sphere and it can be seen that the contour of the patch warps correctly with the contour of the sphere throughout the tracking process. In the final images of the sequence (c) and (d), the pixels which are on the edge of the sphere begin to become occluded. In this case the rigid geometric structure defined within the quadrifocal estimation naturally wraps the edge of the sphere back onto itself, i.e. this is only feasible geometric solution to the estimation problem. Of course, if a robust estimator is not used then tracking fails but with the M-estimator these pixels are no longer used for estimation and are rejected. A test was, however, performed when there was no occlusion which showed that the robust estimator required 10% more iterations to converge than without (or was not as precise with the same number of iterations). There is therefore a compromise to be made between efficiency, precision and robustness. In (e) one can see the estimated trajectory of the camera pair. In (f) it can be seen that even if the root mean square (RMS) error is very small, there is an increase in interpolation error as the camera gets further away from the reference image.

6.2. Visual Odometry on Real Sequences

The algorithm was tested on several real full-scale sequences of urban scenes from different streets in Versailles, France, as can be seen in Figures 8 and 9. Radial distortion has been removed from the images before processing. In these sequences the stereo images are of size 760 × 578 and acquisition is performed at around 15 Hz, the cameras are not rectified and have a baseline of approximately 1 m. In the experiments the bottom half of the image was removed since it contains only the road. The Versailles sequence is available as Extensions 1 and 2. Some further video demonstrations are available online at the authors websites.

The sequence shown in Figure 8 is that of a relatively straight road. The distance travelled by the car has been measured using road markings in the images and satellite views with a precision of 2.9 cm/pixel for the Versailles region. The path length measured by both Google Earth and the tracker was about 440 m. It is difficult to register the satellite image with the projection of the trajectory since no three non-collinear points were available and the best that can be said is that they are approximately the same absolute length (ignoring tilt of the cameras and the incline of the road). Throughout the sequence several moving vehicles pass in front of the cameras and at one stage a car is overtaken.

The sequence shown in Figure 9, is particularly illustrative since a full loop of the roundabout was performed. In particular, this enables the drift to be measured at the crossing point in the trajectory. The drift was measured by comparing the integrated odometry (from the trajectory estimated going around the roundabout) with the pose estimated between images that are close at the closure of the loop (see Figure 10). Although there is no real loop intersection, a local pose estimate is more accurate than integrating many more poses around the entire loop. When there is no intersection, however, there is a non-negligible distance between the images making tracking more difficult and a sufficient overlap is required between the views. This experiment therefore also gives an idea of the convergence domain for the proposed technique (even if this depends on the scene being viewed). In particular, in this experiment it was possible to track directly (no pose initialization) between the initial loop images and the second pass for several image pairs with a maximum estimated distance of 8 m which is much larger that the maximum inter-frame distance reported earlier. As an example, consider the loop closure from the 22nd to the 471st where images the drift estimate was computed from

$$T_{\text{drift}} = \text{inv}(T_{22}) * T_{\text{loopclose}} * T_{471}, \tag{28}$$

which corresponds to an error pose of

$$(-81.0\text{ cm}, -60.8\text{ cm}, 92.9\text{ cm}, 0.1^\circ, 0.2^\circ, 0.2^\circ).$$

This result gives a RMS error at the crossing point of 137.5 cm leading to a drift over the 220 m travelled of approximately 0.6%.

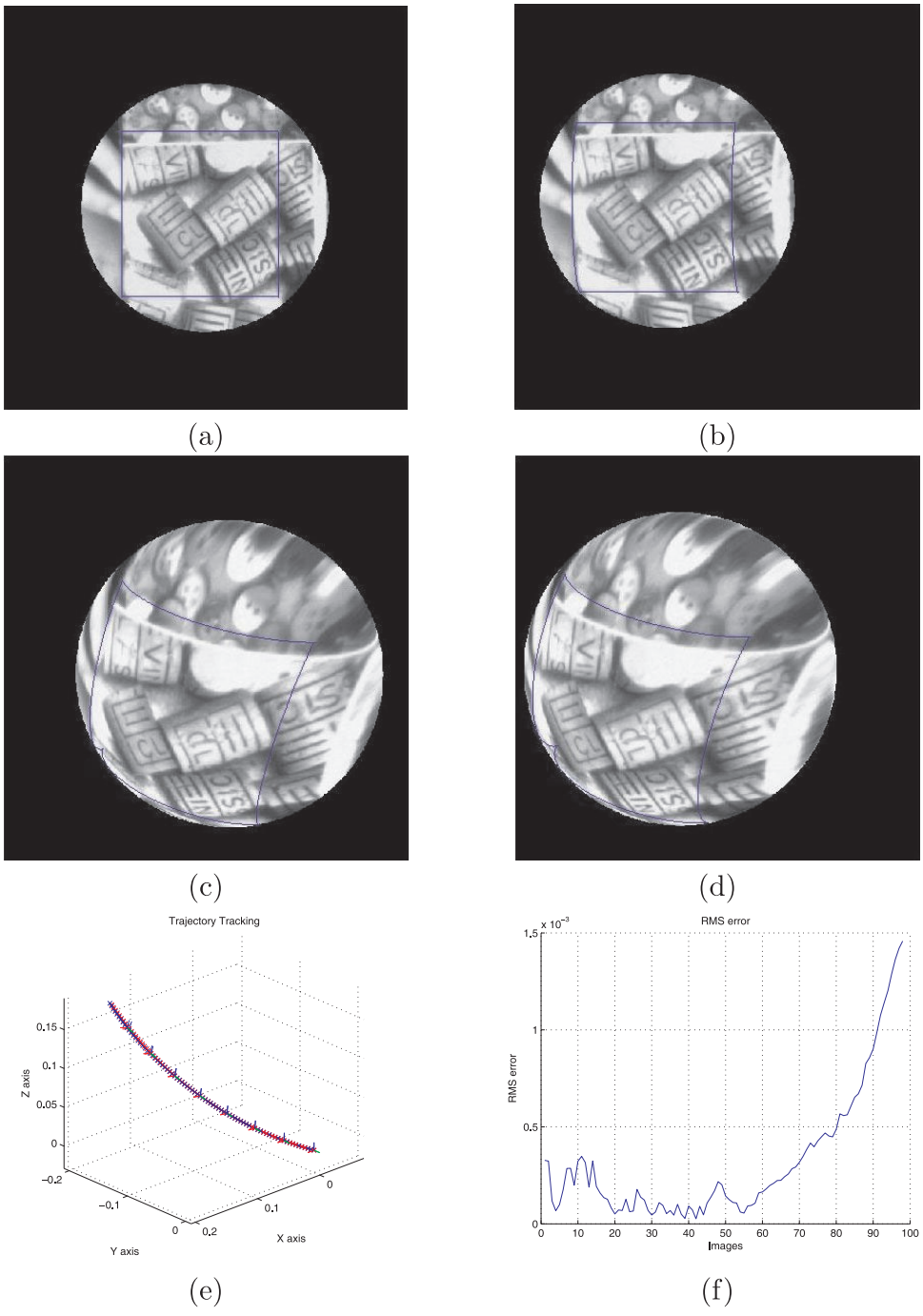


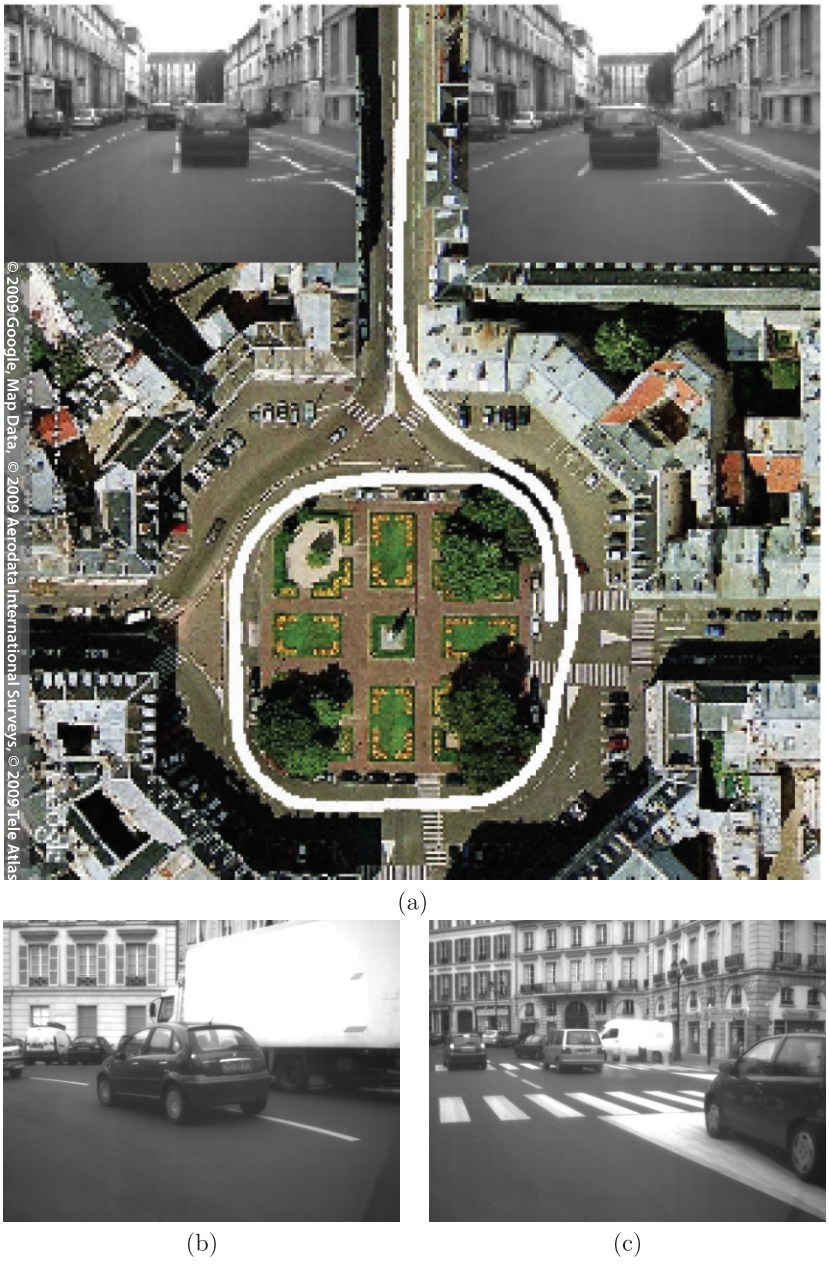
Fig. 7. Simulation of tracking a patch (outlined by the contour) on a 3D sphere: (a), (b) initial left and right images; (c), (d) final left and right images; (e) the estimated trajectory of the sphere; (f) the RMS error between the estimated trajectory and the true trajectory showing drift due to distance from the reference image and less information from the self-occlusion. This simulator was also used to test the computational efficiency of the algorithm versus the size of the image (see Figure 13).

In the case of large-scale scenes such as this one it was necessary to detect and update the reference image periodically when it was no longer visible or too approximate. This update was detected by putting a maximum threshold on the error norm along with a threshold on the median absolute deviation (a robust measure of the standard deviation) and re-

Fig. 8. Trajectory tracking along a road in Versailles: the trajectory shown in white has been superimposed on a satellite image. An typical stereo image is shown at the top.

initializing the dense correspondences. Owing to the highly redundant amount of data, the robust estimator was able to successfully reject pedestrians and moving cars from the estimation process. It can be noted, however, that all static information available was used to estimate the pose (including the parked cars) therefore leading to a very precise result with minimal drift over large displacements.

The experiment shown in Figure 11 is that of the LAAS robotic airship with a wide baseline stereo pair (approximately 2.1 m) looking down towards the ground. The images are taken by two Sony XCD-700 cameras of size 1,024 × 768 with square pixels of 6.25 μm in size. The airship is attached to a string and the images are taken at an average altitude of approximately 25 m. A full loop is performed around a parking lot. This application demonstrates the full potential of the approach since the full six-degree-of-freedom visual odometry is necessary in the aerial domain so as to provide the position of the airship in 3D space. It is important to note that due to the fact that the cameras are facing downwards, the movement observed within the images is very rapid and rough. Furthermore, the balloon is also quite unstable and subject to wind variation. Consequently, the use of predictive filtering in such a context is most often worse than not having a filter. The results in Figure 11 have therefore been obtained without any filter. Unfortunately, this means that there are large inter-frame movements and the visual odometry requires some time to converge (many more iterations). Nevertheless, with the combined use of multi-resolution tracking and the selection of the strongest image gradients the visual odometry computation is still able to be maintained in real time. The results given here have been obtained with an image resolution divided by six and only the 50,000 strongest gradients have been used. Finally, it can also be noted that some local minima were observed at full resolution when the car parking spaces were displaced by the movement of the camera such that the current image of the cars where shifted into an adjacent car park with respect to the reference image. This is due to the fact that the symmetry of the visual information corresponds to the movement of the cam-



(a)

(b) (c)

Fig. 9. Trajectory tracking around a roundabout in Versailles. (a) The trajectory shown in white has been superimposed on a satellite image and it can be seen visually that the trajectory aligns with the four corners of the roundabout (four points are required to estimate the pose). The length of the path is approximately 392 m taken in 698 images. The maximum inter-frame displacement was 1.78 m and the maximum inter-frame rotation was 2.23°. (b), (c) Several occlusions which occurred during the sequence and image 300 and 366, respectively (on the right-hand side of the roundabout).

era. This problem, however, was also avoided by considering a different resolution in the pyramid.

In Figure 12 another experiment is given for a large outdoor sequence with rough natural terrain. Here the images are rectified of size 512 × 384 and captured at about 10 Hz. In this case RTK GPS ground truth data was available that is accurate to within several centimeters. It can be seen here that there is a systematic drift in the direction normal to the Earth surface along with drift that depends on the direction of rotation. This could be due to various factors including camera

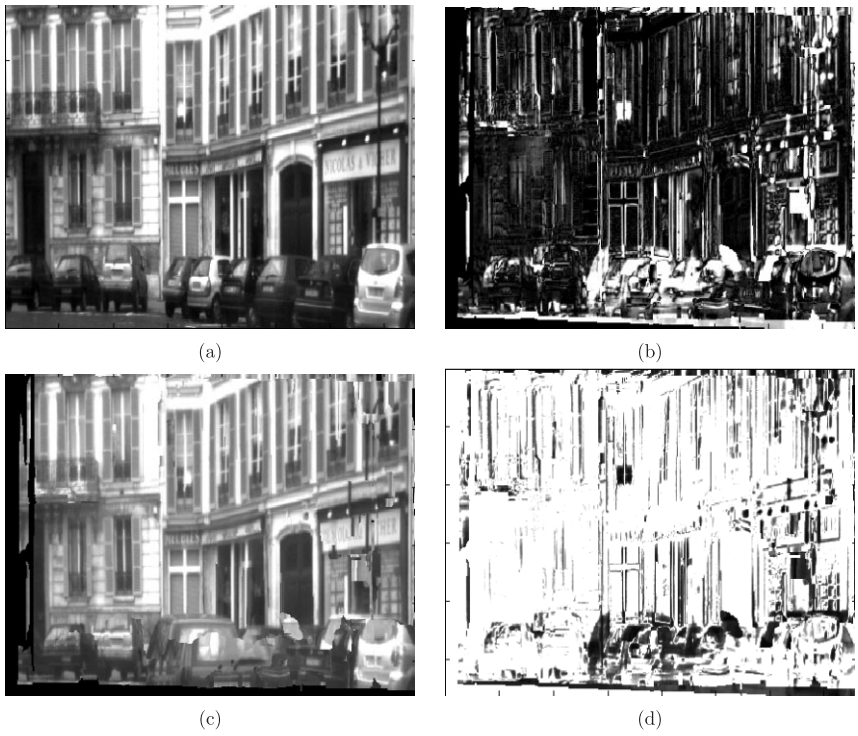


Fig. 10. Loop closing between the 22nd and the 471st images of the Versailles roundabout sequence. (a) The left reference image (the 471st image). (b) The final image error for the left camera. (c) The 22nd image warped to the position of the reference image. (d) The estimated rejection weights. It can be seen that there is a global change in illumination between the images leading to a higher final error and more outliers than in the incremental tracking case.

calibration errors (extrinsic or intrinsic) or errors in the dense matching between the left and the right images. Towards the end of the trajectory a large deviation can be seen due to a very large rotation on the last corner leading to a tracking failure. This could be dealt with by integrating inertial sensors into the framework, as has been done by Konolige and Agrawal (2008).

6.3. Computational Requirements

An optimized version has been implemented which is capable of running in real time, at 60 Hz on a Dual Core T7500 2.2 GHz laptop system with 2 GB of RAM, for images reduced to a resolution of size 100 × 100 (most likely a modern desktop system will produce even better results). In Figure 13, six image resolutions from the full image of size 500 × 500 down to the lowest resolution of size 83 × 83 were considered. It can be seen that the highest resolution runs at 3 Hz and decreases exponentially towards lower precision with resolution 1/2 at 10 Hz and resolution 1/3 at 21 Hz. Each core was used in parallel to handle the data of the left and right images. It can be seen that the computation time varies depending on the size of the image used, the precision required and the magnitude of the displacements considered. In the timings given in Figure 13, the cost of the image display is not counted and only a read from disk is being performed. Gaussian noise with a standard deviation of five gray levels was added to the image measurements (both reference and current) so as to simulate similar error statistics to the real sequences treated earlier. The iterative estimator was limited to five iterations so that a comparison could be made between computational time and precision.

When considering the real image sequences, the robustness to outliers is reduced by using less information, however, it can be noted that there is only a small difference in precision between the full and reduced images. With the Versailles sequence (360 m total distance) there was only about 0.004% drift in translation and 0.03% drift in rotation when comparing the full resolution with 1/3 of the resolution.

Furthermore, a real-time dense correspondence algorithm was also developed and run on the laptop Nvidia QuadroFX 570m GPU. This algorithm is based on a simple local window SSD correlation approach. It was run with a window size of 13 × 13 and 0.5 sub-pixel matching was performed. It is also able to run in parallel to the real-time dual core tracking. For a full image of size 500 × 500 this algorithm requires only 124 ms per image to perform dense matching (computed over an average of 100 images) leading to a computation rate of

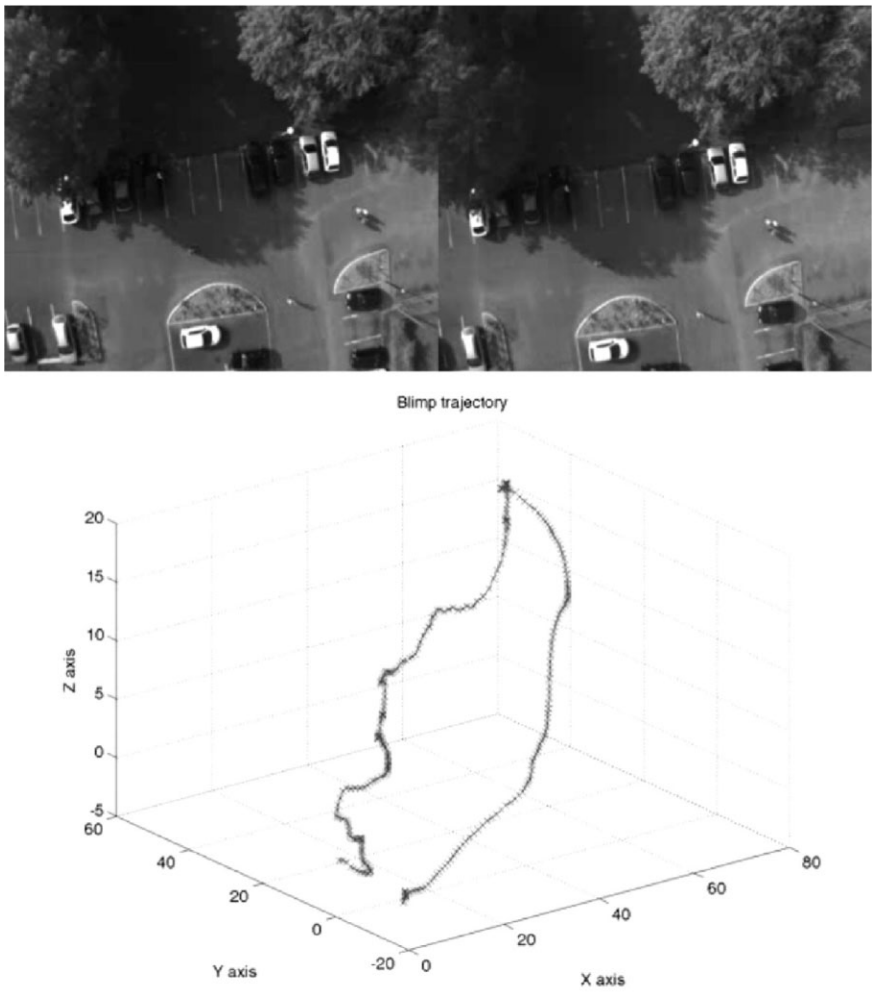


Fig. 11. Trajectory tracking from the LAAS blimp sequence. The trajectory is shown underneath and a pair of typical stereo images is given above.

about 4 Hz for symmetric sub-pixel matching. Since the reference images are kept for a certain duration this computational burden does not slow down the overall tracking frame rates given previously.

The real-time implementation of this approach has required much optimization and various techniques are proposed. In terms of dense correspondence it is possible to:

1. perform dense correspondences online using the GPU making it possible to handle environmental changes over time easily (as for the new results given in this paper);
2. use an off-line training sequence to obtain a set of dense corresponding reference image-pairs for use on-line (as tested and reported by Comport et al. (2007)); this makes it possible to use more accurate but computationally expensive matching techniques, however, it is less robust to changes in the environment (i.e. night or day, vehicles that have moved, etc.).

In terms of the visual odometry it is possible to:

1. contract the quadrifocal transfer function into constant and non-constant components;
2. decompose the Jacobian as outlined in Appendix A;
3. parallelize the computation for single input multiple data (SIMD) architectures such as multi-core and GPU processors;
4. improve the algorithms by multi-resolution and feature selection (i.e. the strongest gradients).

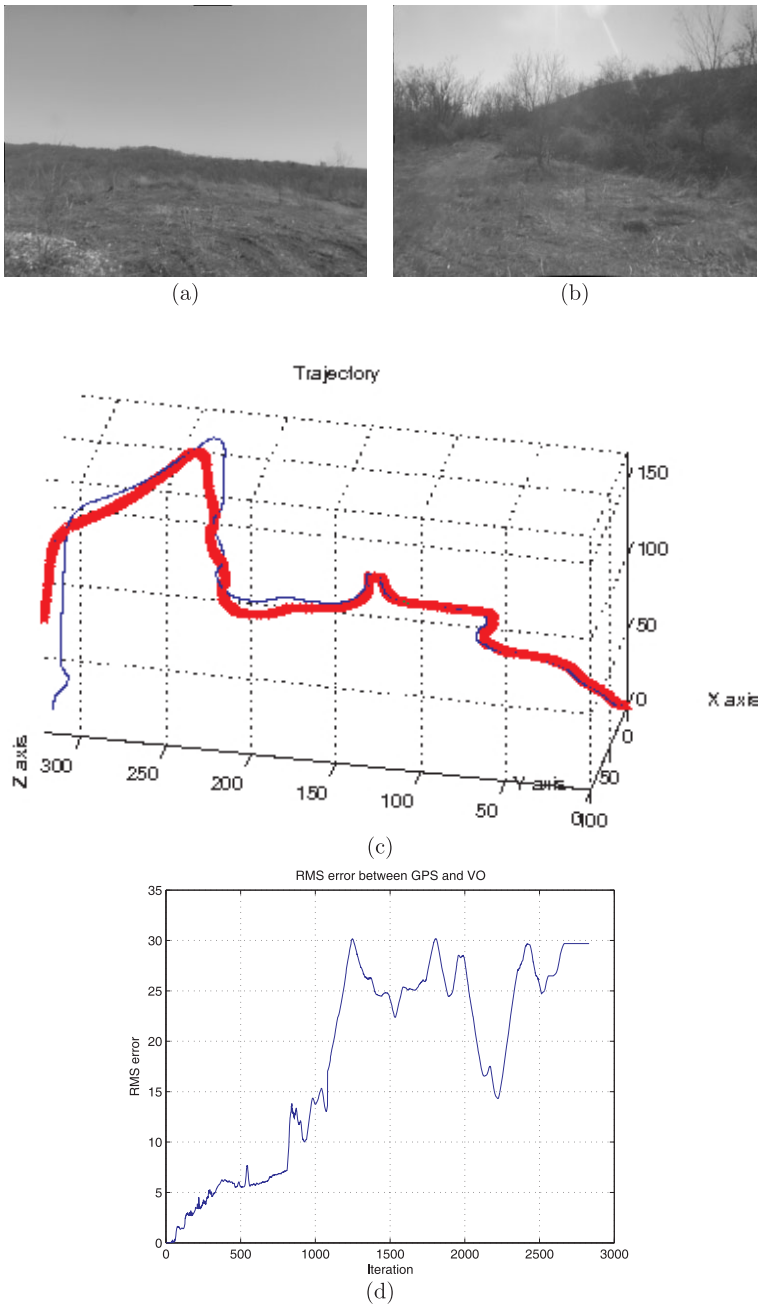


Fig. 12. The outdoor dunes sequence with GPS ground truth taken from a ground vehicle. (a),(b) Two images from this difficult sequence. (c) GPS ground truth data in blue and the visual odometry estimation in red. (d) The RMS error in meters between the GPS and visual sensors: it can be seen that there is significantly more drift when there are large rotations anti-clockwise from iterations 1,081 to 1,244 and clockwise from iteration 1,996 to 2,121.

7. Conclusions and Future Work

The real-time quadrifocal tracking methodology described in this paper has been shown to be very efficient, accurate (very small drift) and robust over a wide range of scenarios. This direct approach is interesting because trajectory estimation is integrated into a single global sensor-based process that does not depend on intermediate level feature extraction nor on region template selection within the image. Furthermore, temporal matching is obtained as the result of the tracking process so no temporal matching is required as in the case of feature-based approaches. In the proposed approach, a compact image-

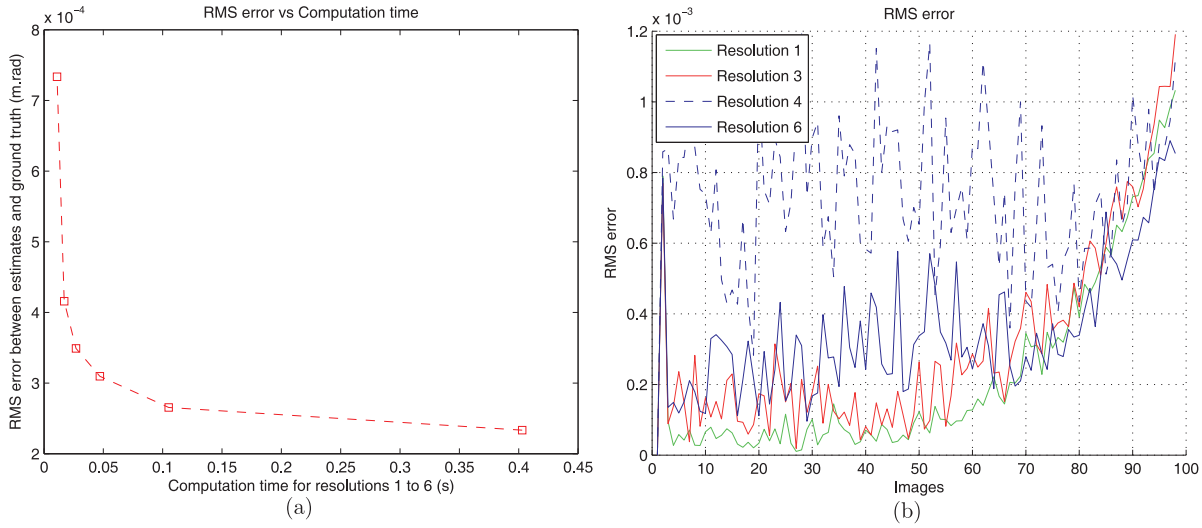


Fig. 13. The scalability of the algorithm depends on the real-time capabilities of the computer architecture and the expected precision of the odometry. In this experiment a simulated sequence of 100 images was generated with a ground truth and the trajectory given in Figure 7 was used. For the purposes of this experiment visual data was made available in the entire image (i.e. no empty regions so as to not bias the results). (a) Shows a plot of the computational time vs the precision of the estimate with values averaged over the entire sequence. Different precisions are obtained by using multiple image resolutions from 1 to 1/6 of a 500×500 image and different computational times are obtained by limiting each estimation to 5 iterations. (b) Shows the RMS error between estimated values and ground truth for different resolutions. It can be seen that as the camera moves further from the reference image the interpolation error begins to outweigh the differences in precision between various image resolutions (at around image 80).

based stereo model of the environment is obtained easily using standard dense stereo correspondence therefore overcoming the difficulty in obtaining an *a priori* 3D model. The robust efficient second-order minimization technique also allows minimization of a highly redundant non-linear function in a precise manner. Indeed the algorithm rejects outliers such as pedestrians, traffic, building occlusions and matching error.

Further work will be devoted to estimating optimal stereo image-based models of the environment by updating the dense correspondences in a simultaneous localization and correspondence style approach. It would also be interesting to further test loop closing procedures and devise strategies to recognize previously seen places within this framework.

Acknowledgement

This study was part of the French national MOBIVIP PREDIT project aimed at autonomous vehicle navigation in urban environments. Part of this work was also carried out at LASMEA, Clermont-Ferrand, France.

Appendix A: Jacobian Computation

A.1. Current Jacobian

The current Jacobian $\mathbf{J}(\mathbf{0})$ can be obtained by taking the derivative of Equation (5) and evaluating at $\mathbf{x} = \mathbf{0}$ as

$$\mathbf{J}(\mathbf{0}) = \left[\nabla_{\mathbf{x}} \mathcal{I}\left(w\left(\mathcal{P}^*, \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})\right)\right)\right]_{\mathbf{x}=\mathbf{0}}. \tag{29}$$

Taking into account property (16) gives

$$\mathbf{J}(\mathbf{0}) = \left[\nabla_{\mathbf{x}} \mathcal{I}\left(w\left(w\left(\mathcal{P}^*, \mathbf{T}(\mathbf{x})\right), \widehat{\mathbf{T}}\right)\right)\right]_{\mathbf{x}=\mathbf{0}}, \tag{30}$$

which can be written as a product of four Jacobians:

$$\mathbf{J}(\mathbf{0}) = \mathbf{J}_{\mathcal{I}} \mathbf{J}_w \mathbf{J}_{\mathbf{T}} \mathbf{J}_{\mathcal{V}}. \tag{31}$$

1. The Jacobian $\mathbf{J}_{\mathcal{I}}$ is of dimension $1 \times 2 * 3 = 6$ and corresponds to the spatial derivative of the pixel intensities for each of the current images warped by the projective transformation $w(\mathbf{z}, \widehat{\mathbf{T}})$

$$\mathbf{J}_{\mathcal{I}} = \left[\nabla_{\mathbf{z}} \mathcal{I}\left(w(\mathbf{z}, \widehat{\mathbf{T}})\right)\right]_{\mathbf{z}=\mathcal{P}^*}. \tag{32}$$

2. The Jacobian $\mathbf{J}_w$ is of dimension $6 \times 2 * 16 = 32$ and corresponds to the derivative of the pixel location with respect to the elements of two homogeneous pose matrices embedded within each of the two trifocal tensor of the warping function

$$\mathbf{J}_w = \left[\nabla_{\mathcal{Z}} w\left(\mathcal{P}^*, \mathcal{Z}\right)\right]_{\mathcal{Z}=\mathbf{T}(\mathbf{0})=\mathbf{I}}. \tag{33}$$

3. The Jacobian $\mathbf{J_T}$ is of dimension $32 \times 6 * 2 = 12$ and can be written as

$$\mathbf{J_T} = \nabla_{\mathbf{x_L},\mathbf{x_R}} \begin{bmatrix} \mathbf{T_L}(\mathbf{x_L}) \\ \\ \mathbf{T_R}(\mathbf{x_R}) \end{bmatrix}_{\mathbf{x_L}=\mathbf{x_R}=0}, \tag{34}$$

where both left and right camera matrices are not expressed in the same reference frame but with respect to their canonical trifocal tensor bases. This Jacobian can be written as

$$\mathbf{J_T} = [\mathbf{a}_{\mathbf{L}1}, \ldots, \mathbf{a}_{\mathbf{L}6}, \mathbf{a}_{\mathbf{R}1}, \ldots, \mathbf{a}_{\mathbf{R}6}], \tag{35}$$

where the vectors $\mathbf{a_i}$ are obtained by reshaping the generator matrices A_i (columns then rows). The generators are the basis of the Lie algebra $\mathfrak{se}$ and can be determined from Equation (4) by selecting a basis x from the twist $\mathbf{x}_i = (\boldsymbol{v}_i, \boldsymbol{\omega}_i)$ (i.e. $\mathbf{v_1} = (1, 0, 0, 0, 0, 0)$) to obtain

$$\mathbf{A}_i = \begin{bmatrix} [\boldsymbol{\omega}_i]_\times & \boldsymbol{v}_i \\ \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{36}$$

4. The Jacobian $\mathbf{J}_\mathcal{V}$ of dimension 12×6 is a pair of adjoint maps that transform the twists from the precedent Jacobian into the same reference frame according to Equation (12). It is chosen to center the two components of $\mathbf{J_T}$, corresponding to the left and right canonical coordinate systems, so that they represent the same minimal set of unknown parameters. This corresponds to the application of the bilinear constraint to the pair of trifocal constraints so as to form a quadrilinear constraint giving

$$\mathbf{J}_\mathcal{V} = \begin{bmatrix} \frac{\partial \mathbf{x_L}}{\partial \mathbf{x}} \\ \\ \frac{\partial \mathbf{x_R}}{\partial \mathbf{x}} \end{bmatrix} = \begin{bmatrix} \mathbf{V_L} \\ \\ \mathbf{V_R} \end{bmatrix}, \tag{37}$$

where the adjoint map is given as

$$\mathcal{V} = \begin{bmatrix} \mathbf{R^c} & \mathbf{t^c}_\times \mathbf{R^c} \\ \\ \mathbf{0}_3 & \mathbf{R^c} \end{bmatrix}, \tag{38}$$

and where $\mathbf{T}^c = (\mathbf{R^c}, \mathbf{t^c})$ is the centering pose given in Equation (11), which maps the current left camera matrix to the stereo center according to Equation (11). Similarly, an adjoint map can be obtained to transform the twist of the right current camera with respect to the right reference camera using $\mathbf{T}'^c$.

Three of the Jacobians, $\mathbf{J}_w$, $\mathbf{J_T}$ and $\mathbf{J}_\mathcal{V}$ are constant and need only be calculated once for the reference image. The Jacobian $\mathbf{J}_\mathcal{I}$ must be calculated at each iteration.

A.2. Reference Jacobian

The reference Jacobian $\mathbf{J}(\tilde{\mathbf{x}})$ can be obtained by taking the derivative of Equation (5) as

$$\mathbf{J}(\mathbf{x}) = \left[\nabla_\mathbf{x} \boldsymbol{\mathcal{I}} \left(w(\boldsymbol{\mathcal{P}}^*, \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})) \right)\right]_{\mathbf{x}=\tilde{\mathbf{x}}}. \tag{39}$$

The group property (16) is reused again here to replace the current image by the reference image. This is achieved by introducing a *true* transformation $\overline{\mathbf{T}}$ corresponding to the solution that is sought after along with its inverse. In this case Equation (39) can be rewritten as

$$\begin{aligned} \mathbf{J}(\mathbf{x}) &= \left[\nabla_\mathbf{x} \boldsymbol{\mathcal{I}} \left(w \left(w(\boldsymbol{\mathcal{P}}^*, \overline{\mathbf{T}}^{-1}\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})), \overline{\mathbf{T}} \right) \right)\right]_{\mathbf{x}=\tilde{\mathbf{x}}}, \\ &= \left[\nabla_\mathbf{x} \boldsymbol{\mathcal{I}}^* \left(w(\boldsymbol{\mathcal{P}}^*, \overline{\mathbf{T}}^{-1}\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})) \right)\right]_{\mathbf{x}=\tilde{\mathbf{x}}}. \end{aligned} \tag{40}$$

The reference Jacobian can now be written as a product of four Jacobians:

$$\mathbf{J}(\tilde{\mathbf{x}}) = \mathbf{J}_{\boldsymbol{\mathcal{I}}^*} \mathbf{J}_{w^*} \mathbf{J}_{\mathbf{T}^*} \mathbf{J}_{\mathbf{V}^*}. \tag{41}$$

1. The Jacobian $\mathbf{J}_{\boldsymbol{\mathcal{I}}^*}$ is of dimension $1 \times 2 * 3 = 6$ and corresponds to the spatial derivative of the pixel intensities for each of the reference images warped by the projective transformation $w(\mathbf{z}, \overline{\mathbf{T}}^{-1}\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}))$:

$$\mathbf{J}_{\boldsymbol{\mathcal{I}}^*} = \left[\nabla_\mathbf{z} \boldsymbol{\mathcal{I}}^* \left(w(\mathbf{z}, \mathbf{I})\right)\right]_{\mathbf{z}=\boldsymbol{\mathcal{P}}^*}, \tag{42}$$

where $\overline{\mathbf{T}}^{-1}\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}) = \mathbf{I}$ at $\mathbf{x} = \tilde{\mathbf{x}}$.

2. The Jacobian $\mathbf{J}_{w^*}$ is of dimension $6 \times 2 * 16 = 32$ and corresponds to the derivative of the pixel location with respect to the elements of two homogeneous pose matrices embedded within each of the two trifocal tensor of the warping function:

$$\mathbf{J}_{w^*} = \left[\nabla_\mathcal{Z} w \left(\boldsymbol{\mathcal{P}}^*, \mathcal{Z}\right)\right]_{\mathcal{Z}=\mathbf{I}} = \mathbf{J}_w. \tag{43}$$

3. The Jacobian $\mathbf{J}_{\mathbf{T}^*}$ depends on the unknown twist $\tilde{\mathbf{x}}$ which is the solution to the estimation problem. However, by using the group properties it can be shown that the following proposition is true (see Appendix B):

$$\mathbf{J}_{\mathbf{T}^*} \mathbf{J}_{\mathcal{V}^*} \mathbf{x} = \mathbf{J_T} \mathbf{J}_\mathcal{V} \mathbf{x}. \tag{44}$$

4. The Jacobian $\mathbf{J}_{\mathbf{V}^*} = \mathbf{J}_\mathcal{V}$ since the same rigid transformation is made.

Appendix B: Proof of the proposition in Equation (44)

Here a small proof is given for the proposition in Equation (44). The proposition is first rewritten as

$$\left[\frac{d(\overline{\mathbf{T}}^{-1}\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}))}{d\mathbf{x}}\right]_{\mathbf{x}=\tilde{\mathbf{x}}} \tilde{\mathbf{x}} = \left[\frac{d\mathbf{T}(\mathbf{x})}{d\mathbf{x}}\right]_{\mathbf{x}=\mathbf{0}} \tilde{\mathbf{x}}. \tag{45}$$

Considering the left-hand side, it is possible to make a substitution of variables for $\mathbf{x} = \tilde{\mathbf{x}} + \mathbf{y}$ to give

$$\mathbf{J}_{\mathbf{T}^*} = \left[\frac{d(\overline{\mathbf{T}}^{-1}\widehat{\mathbf{T}}\mathbf{T}(\tilde{\mathbf{x}}+\mathbf{y}))}{d\mathbf{y}}\right]_{\mathbf{x}=\mathbf{0}} \frac{d\mathbf{y}}{d\mathbf{x}}\tilde{\mathbf{x}}, \tag{46}$$

where using the group properties $\mathbf{T}(\tilde{\mathbf{x}})\mathbf{T}(\mathbf{y}) = \mathbf{T}(\tilde{\mathbf{x}}+\mathbf{y})$ and assuming that the "true" pose $\overline{\mathbf{T}} \approx \widehat{\mathbf{T}}\mathbf{T}(\tilde{\mathbf{x}})$ gives

$$\mathbf{J}_{\mathbf{T}^*} = \left[\frac{d\mathbf{T}(\mathbf{y})}{d\mathbf{y}}\right]_{\mathbf{y}=\mathbf{0}} \tilde{\mathbf{x}}, \tag{47}$$

which gives the same $\tilde{\mathbf{x}}$ as in the right-hand side of Equation (45) for all $\mathbf{y} = \mathbf{x} - \tilde{\mathbf{x}}$.

Appendix C: Robust M-estimation

In this appendix we give a brief overview for the calculation of weights for each image feature. The weights w_i, which represent the different elements of the $\mathbf{D}$ matrix and reflect the confidence of each feature, are usually given by (Huber 1981) as

$$w_i = \frac{\psi(\delta_i/\sigma)}{\delta_i/\sigma}, \tag{48}$$

where

$$\psi(\delta_i/\sigma) = \frac{\partial\rho(\delta_i/\sigma)}{\partial\mathbf{r}},$$

where ψ is the influence function, and δ_i is the normalized residual given by $\delta_i = \Delta_i - \text{Med}(\Delta)$ (where Med(Δ) is the median operator).

Of the various loss and corresponding influence functions that exist in the literature Tukey's hard re-descending function is considered. Tukey's function completely rejects outliers and gives them a zero weight. This is of interest in tracking applications so that a detected outlier has no effect on the virtual camera motion and does not cost computational effort uselessly. This influence function is given by

$$\psi(u) = \begin{cases} u(C^2-u^2)^2 & \text{, if } |u| \le C, \\ 0 & \text{, otherwise,} \end{cases} \tag{49}$$

where the proportionality factor for Tukey's function is $C = 4.6851$ and represents 95% efficiency in the case of Gaussian noise.

In order to obtain a robust objective function, a value describing the certainty of the measures is required. The scale σ or the estimated standard deviation of the inlier data and is a critical value that can impact heavily on the efficiency of the method. This factor varies significantly during convergence, so it is estimated iteratively using the median absolute deviation:

$$\widehat{\sigma} = \frac{1}{\Phi^{-1}(0.75)}\text{Med}_i(|\delta_i - \text{Med}_j(\delta_j)|), \tag{50}$$

where $\Phi(\ )$ is the cumulative normal distribution function and $1/\Phi^{-1}(0.75) = 1.48$ represents one standard deviation of the normal distribution.

The introduction of the weighting matrix $\mathbf{D}$ into the minimization scheme in Section 4 is achieved via an iteratively re-weighted least-squares implementation.

Appendix D: Index to Multimedia Extensions

The multimedia extension page is found at http://www.ijrr.org

Table of Multimedia Extensions

Extension	Type	Description
1	Video	A video showing the estimated trajectory overlaid in blue on a satellite image from Google Earth for the Versailles roundabout sequence. The input stereo images are also overlaid on the video sequence.
2	Video	A video showing the warped current images (for a given reference image), the weights that were computed and the input stereo images used in the estimation of the trajectory. This data is also from the Versailles roundabout sequence. One can see that the reference images are held for a certain duration and then updated according to the robust statistical measures of the error standard deviation and the mean.

References

Avidan, S. and Shashua, A. (2001). Threading fundamental matrices. *Pattern Analysis and Machine Intelligence*, **23**(1): 73–77.

Baker, S. and Matthews, I. (2001). Equivalence and efficiency of image alignment algorithms. *Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition*.

Benhimane, S. and Malis, E. (2004). Real-time image-based tracking of planes using efficient second-order minimization. *IEEE International Conference on Intelligent Robots Systems*, Sendai, Japan.

Burt, P. (1984). The pyramid as structure for efficient computation. *Multiresolution Image Processing and Analysis*. Berlin, Springer, pp. 6–35.

Chiuso, A., Favaro, P., Jin, H. and Soatto, S. (2002). Structure from motion causally integrated over time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(4): 523–535.

Comport, A., Malis, E. and Rives, P. (2007). Accurate quadrifocal tracking for robust 3D visual odometry. *IEEE International Conference on Robotics and Automation*, Rome, Italy.

Comport, A., Marchand, E. and Chaumette, F. (2006a). Statistically robust 2D visual servoing. *IEEE Transactions on Robotics*, **22**(2): 415–421.

Comport, A., Marchand, E., Pressigout, M. and Chaumette, F. (2006b). Real-time markerless tracking for augmented reality: the virtual visual servoing framework. *IEEE Transactions on Visualization and Computer Graphics*, **12**(4): 615–628.

Davison, A. J. and Murray, D. W. (2002). Simultaneous localisation and map-building using active vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**: 865–880.

Faugeras, O. and Mourrain, B. (1995). On the geometry and algebra of the point and line correspondences between *n* images. *IEEE International Conference on Computer Vision*, Cambridge, MA, pp. 951–956.

Hager, G. and Belhumeur, P. (1998). Efficient region tracking with parametric models of geometry and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(10): 1025–1039.

Hartley, R. (1995). Multilinear relationships between coordinates of corresponding image points and lines. *Proceedings of the Sophus Lie Symposium*, Nordjordeid, Norway.

Hartley, R. and Zisserman, A. (2001). *Multiple View Geometry in computer vision*. Cambridge, Cambridge University Press.

Heckbert, P. (1989). *Fundamentals of Texture Mapping and Image Warping*. Master's thesis, CS Division, U.C. Berkeley.

Heyden, A. and Astrom, K. (1997). Algebraic properties of multilinear constraints. *Mathematical Methods in the Applied Sciences*, **20**(13): 1135–1162.

Howard, A. (1983). Scale-space filtering. *8th International Joint Conference on Artificiel Intelligence*, Karlsruhe, Germany, pp. 1019–1022.

Howard, A. (2008). Real-time stereo visual odometry for autonomous ground vehicles. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nice, France.

Huber, P.-J. (1981). *Robust Statistics*. New York, Wiley.

Irani, M. and Anandan, P. (2000). About direct methods. *IEEE International Conference on Computer Vision: Proceedings of the International Workshop on Vision Algorithms*. London, Springer, pp. 267–277.

Konolige, K. and Agrawal, M. (2008). Frameslam: from bundle adjustment to realtime visual mapping. *IEEE Transactions on Robotics*, **24**(5): 1066–1077.

Lucas, B. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *International Joint Conference on Artificial Intelligence*, Vancouver, Canada, Vol. 2, pp. 674–679.

Malis, E. (2004). Improving vision-based control using efficient second-order minimization techniques. *IEEE International Conference on Robotics and Automation*, New Orleans, LA, Vol. 2, pp. 1843–1848.

Mouragnon, E., Lhuillier, M., Dhome, M., Dekeyser, F. and Sayd, P. (2006). Real-time localization and 3D reconstruction. *IEEE Conference of Vision and Pattern Recognition*, New York.

Nistér, D., Naroditsky, O. and Bergen, J. (2004). Visual odometry. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, Vol. 1, pp. 652–659.

Nister, D., Naroditsky, O. and Bergen, J. (2006). Visual odometry for ground vehicle applications. *Journal of Field Robotics*, **23**: 2006.

Odobez, J.-M. and Bouthemy, P. (1995). Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, **6**(4): 348–365.

Ogale, A. and Aloimonos, Y. (2005). Shape and the stereo correspondence problem. *International Journal of Computer Vision*, **65**: 147–162.

Scharstein, D., Szeliski, R. and Zabih, R. (2001). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Proceedings of the IEEE Workshop on Stereo and Multi-Baseline Vision*, Kauai, HI.

Shashua, A. and Wolf, L. (2000). On the structure and properties of the quadrifocal tensor. *European Conference on Computer Vision*, pp. 710–724.

Silveira, G., Malis, E. and Rives, P. (2008). An efficient direct approach to visual SLAM. *IEEE Transactions on Robotics*, **20**(5): 969–979.

Simond, N. and Rives, P. (2008). What can be done with an embedded stereo-rig in urban environments? *Robotics and Autonomous Systems*, **56**: 777–789.

Triggs, B. (1995). The geometry of projective reconstruction. I: Matching constraints and the joint image. *IEEE International Conference on Computer Vision*, Cambridge, MA, pp. 338–343.

van der Mark, W. and Gavrila, D. (2006). Real-time dense stereo for intelligent vehicles. *IEEE Transactions on Intelligent Transportation Systems*, **7**(1): 38–50.

Zhang, Z. and Faugeras, O. (1992). Three dimensional motion computation and object segmentation in a long sequence of stereo frames. *International Journal of Computer Vision*, **7**(3): 211–241.