Accurate Quadrifocal Tracking for Robust 3D Visual Odometry

A.I. Comport, E. Malis and P. Rives

***Abstract*— This paper describes a new image-based approach to tracking the 6dof trajectory of a stereo camera pair using a corresponding reference image pairs instead of explicit 3D feature reconstruction of the scene. A dense minimisation approach is employed which directly uses all grey-scale 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.**

I. INTRODUCTION

This study is part the MOBIVIP project aimed at autonomous vehicle navigation in urban environments. Here the core issue is 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 [17], [12]. It is clear that pose estimation and visual tracking are also important in many applications including robotics, augmented reality, medical imaging, etc...

Model-based techniques have shown that 3D CAD models are essential for robust, accurate and efficient 3D motion estimation [6], 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 urban environments.

Alternative techniques propose to perform 3D structure and motion estimation online. Among this class, visual simultaneous localisation and mapping approaches [7], [4] are 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 (due to approximate non-iterative estimation). In [13] stereo and monocular visual odometry approaches are proposed based on a combination of feature extraction, matching, tracking, triangulation, RANSAC pose estimation and iterative refinement. In [12], a similar monocular technique is proposed but drift is minimised using a local bundle adjustment technique.

Feature based methods (eg. [7], [4], [13], [12]) all rely on an intermediary estimation processes 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 error and accumulate drift. To eliminate drift these approaches resort to techniques such as local bundle adjustment or SLAM.

Appearance and optical flow based techniques, on the other hand, are image-based and minimise an error directly based on the image measurements. Unfortunately, they are often only monocular and make heavy assumptions about the nature of the structure within the scene or the camera model. For example in [8] an affine camera model is assumed and in [2] and [3] planar homography models are assumed. In this way the perspective effects or the effects of non-planar 3D objects are not considered and tracking fails easily under large movements. Of course many papers avoid the problems of monocular algorithms (i.e. scale factor, initialisation, observability, etc.) by using multi-view constraints and a multitude of work exist on multiview-geometry (see [9] and ref. therein). However, to our knowledge no work has been done on deriving an efficient region-based tracker as in [3] using stereo warping and novel view synthesis as in [1].

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 suitable for initialization procedures. *Full-scale non-linear optimisation techniques* (e.g., [8], [2], [6]) 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 accuracy, however, they may be subject to local minima and, worse, divergence. In this paper an efficient second-order approach [11], [3] is employed which improves efficiency and helps to avoid any local minima.

The technique proposed in this paper provides a generic 3D visual odometry technique which is able to accurately handle large scale scenes efficiently whilst avoiding error prone feature extraction. This is achieved by defining a quadrifocal warping function which closes a non-linear iterative estimation loop directly with the image. A set of key reference image-pairs are used to initialize tracking locally around the reference positions. These reference pairs provide a calibrated set of highly redundant dense correspondences to perform tracking and pose estimation. As will be shown, this leads to very impressive results in real-scenes with occlusions, large inter-frame displacements, and very little

This work was supported by the MOBIVIP PREDIT project.
The authors are with INRIA, Sophia-Antipolis, France.
name.surname@sophia.inria.fr

drift over very long sequences of images.

II. Trajectory Estimation

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

Since our final 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 $\mathbf{p}=(u,v)$ and $\mathbf{p}'=(u',v')$ are 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. Similarly $\boldsymbol{\mathcal{P}}^*=\{\mathbf{p},\mathbf{p}'\}$ are stereo image correspondences from the reference template pair.

$\boldsymbol{\mathcal{I}}$ 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. Any set of corresponding pixels from the reference image-pair are considered as a reference template, denoted by $\mathcal{R}^*=\{\{\mathbf{p}^*,\mathbf{p}^{*\prime}\}_1,\{\mathbf{p}^*,\mathbf{p}^{*\prime}\}_2,\ldots,\{\mathbf{p}^*,\mathbf{p}^{*\prime}\}_n\}$ where n is the number of corresponding point pairs in the stereo images.

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 III. $\mathbf{K}$ and $\mathbf{K}'$ contain the intrinsic calibration parameters for the left and right cameras respectively. $\mathbf{T}'=(\mathbf{R}',\mathbf{t}')\in\mathbb{SE}(3)$ is the homogeneous matrix of the extrinsic camera pose of the right camera w.r.t. the left and $\overline{\mathbf{T}}=(\bar{\mathbf{R}},\bar{\mathbf{t}})\in\mathbb{SE}(3)$ 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:

$$\boldsymbol{\mathcal{I}}^*(\boldsymbol{\mathcal{P}}^*)=\boldsymbol{\mathcal{I}}\big(w(\boldsymbol{\mathcal{P}}^*;\overline{\mathbf{T}}),t\big),\quad\forall\boldsymbol{\mathcal{P}}^*\in\mathcal{R}^*. \tag{1}$$

where $\overline{\mathbf{T}}$ is the true pose.

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 it is supposed that $\exists\tilde{\mathbf{x}}:\mathbf{T}(\tilde{\mathbf{x}})\widehat{\mathbf{T}}=\overline{\mathbf{T}}$. The estimate is updated by a homogeneous transformation $\widehat{\mathbf{T}}\leftarrow\mathbf{T}(\mathbf{x})\widehat{\mathbf{T}}$.

The unknown parameters $\mathbf{x}\in\mathbb{R}^6$ are defined as:

$$\mathbf{x}=\int_0^1(\boldsymbol{\omega},\boldsymbol{\upsilon})dt\in se(3), \tag{2}$$

which is the integral of a constant velocity twist which produces a pose $\mathbf{T}$. The pose and the twist are related via the exponential map as $\mathbf{T}=e^{[\mathbf{x}]_\wedge}$ with the operator $[.]_\wedge$ as:

$$[\mathbf{x}]_\wedge=\begin{bmatrix}[\boldsymbol{\omega}]_\times & \boldsymbol{\upsilon}\\ \mathbf{0} & \mathbf{0}\end{bmatrix},$$

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

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

$$C(\mathbf{x})=\sum_{\boldsymbol{\mathcal{P}}^*\in\mathcal{R}^*}\Big(\boldsymbol{\mathcal{I}}\Big(w(\boldsymbol{\mathcal{P}}^*;\mathbf{T}(\mathbf{x})\widehat{\mathbf{T}}\Big)-\boldsymbol{\mathcal{I}}^*\left(\boldsymbol{\mathcal{P}}^*\right)\Big)^2. \tag{3}$$

This function is minimised using the robust, efficient and precise second order minimisation procedure detailed in Section IV.

III. Novel view synthesis and warping

The geometric configuration of the stereo pair is based on the paradigm that four views of a scene satisfy quadrifocal constraints. Thus given a reference stereo view in correspondence 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 view-pairs in a sequence of images.

A. Quadrifocal Geometry

A point $\mathbf{P}\in\mathbb{R}^3$ in 3D 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 $\bar{\mathbf{p}}=\mathbf{MP}$ so that $\bar{\mathbf{p}}=(u,v,1)^\top$ is the homogeneous pixel vector (see Figure 1).

In [16] it is shown that the quadrifocal tensor can be decomposed into two trifocal tensors. In this way the geometry between two stereo pairs is 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 as : $\mathbf{M}'=\mathbf{K}'[\mathbf{R}'|\mathbf{t}']$ 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. This is done symmetrically by defining the *world origin* as the geodesic center between the two reference cameras. To do this the extrinsic parameters must be separated into two distinct poses with respect to the center as:

$$\mathbf{T}^{\mathbf{c}}=e^{(\log(\mathbf{T}'))/2}\quad\text{and}\quad\mathbf{T}^{\mathbf{c}\prime}=\mathbf{T}^{\mathbf{c}}\mathbf{T}'^{-1}, \tag{4}$$

where e and $\log$ are the matrix exponential and logarithm.

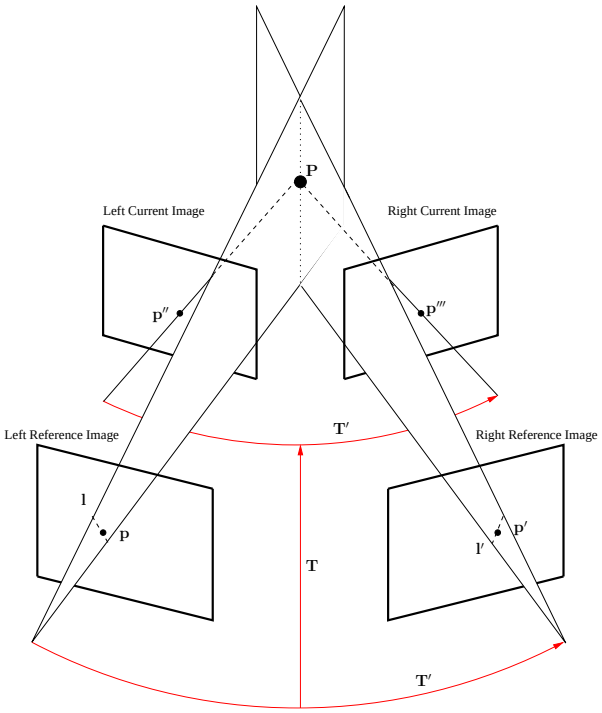


Fig. 1. The quadrifocal geometry of a stereo pair at two subsequent time instants. Two points $\mathbf{p}$ and $\mathbf{p}'$ are initialised only once at the beginning of the tracking process to be in correspondence. The central pose $\mathbf{T}$ is estimated via a non-linear warping function which warps 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.

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

$$\mathbf{T}'' = \mathbf{T}^{\mathbf{c}\,-1}\widetilde{\mathbf{T}}\mathbf{T}^{\mathbf{c}}, \tag{5}$$

where $\widetilde{\mathbf{T}}$ is the unknown pose to be estimated.

B. Quadrifocal warping

The quadrifocal warping function $w(\boldsymbol{\mathcal{P}}^*; \overline{\mathbf{T}})$ from (3) can now be considered to be composed of a trifocal tensor for both left and right images. Even though the warping of each image is carried out separately using two trifocal tensors, these functions will be minimized simultaneously in section IV so that the quadrifocal constraints are held. 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 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.

The trifocal tensor $\mathcal{T}$ is a third order tensor represented by a homogeneous $3 \times 3 \times 3$ array of elements. The calibrated trifocal tensor is given as:

$$\mathcal{T}_i^{jk} = \mathbf{k'}_m^j \mathbf{t}'^m \mathbf{k''}_i^n \mathbf{r}''^o_n \mathbf{k}^{-1}{}_o^k - \mathbf{k''}_m^k \mathbf{t}''^m \mathbf{k'}_i^n \mathbf{r}'^o_n \mathbf{k}^{-1}{}^j,$$

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. $\mathbf{k}$ and $\mathbf{k}'$ are the intrinsic calibration components of the left and right camera matrices respectively. Note that $\mathbf{k}'' = \mathbf{k}$ or $\mathbf{k}'' = \mathbf{k}'$ 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.

The stereo warping operator is then given by:

$$\begin{bmatrix} \mathbf{p}''^k \\ \mathbf{p}'''^n \end{bmatrix} = \begin{bmatrix} \mathbf{p}^i \mathbf{l}'_j \mathcal{T}_i^{jk} \\ \mathbf{p}'^l \mathbf{l}_m \mathcal{T}_l^{mn} \end{bmatrix}, \tag{6}$$

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 similarly to [1] to be the diagonal line $(-1, -1, u+v)$ coincident with the point (u, v).

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

1) The identity map:

$$w(\boldsymbol{\mathcal{P}}^*; \mathbf{I}) = \boldsymbol{\mathcal{P}}^*, \quad \forall \boldsymbol{\mathcal{P}}^* \in \mathbb{R}^4, \tag{7}$$

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

$$w(w(\boldsymbol{\mathcal{P}}^*, \mathbf{T_1}), \mathbf{T_2}) = w(\boldsymbol{\mathcal{P}}^*, \mathbf{T_1}\mathbf{T_2}) \quad \forall \boldsymbol{\mathcal{P}}^* \in \mathbb{R}^4. \tag{8}$$

IV. Robust second-order minimisation

The aim now is to minimise the objective criterion defined previously (3) in an accurate and robust manner. The robust objective function therefore becomes:

$$O(\mathbf{x}) = \sum_{\boldsymbol{\mathcal{P}}^* \in \mathcal{R}^*} \rho\left(\boldsymbol{\mathcal{I}}(w(\boldsymbol{\mathcal{P}}^*; \mathbf{T}(\mathbf{x})\widehat{\mathbf{T}}) - \boldsymbol{\mathcal{I}}^*(\boldsymbol{\mathcal{P}}^*)\right), \quad (9)$$

where $\rho(u)$ is a robust function [10] that grows sub-quadratically and is monotonically non-decreasing with increasing $|u|$ (see [5]).

Since this is a non-linear function of the unknown pose parameters an iterative minimisation procedure is employed. The robust objective function is minimized by: $\nabla_{\mathbf{x}} O(\mathbf{x})|_{\mathbf{x}=\mathbf{x}} = 0$, where $\nabla_{\mathbf{x}}$ is the gradient operator with respect to the unknown parameters (2) and there exists a stationary point $\mathbf{x} = \tilde{\mathbf{x}}$ which is the global minimum of the cost function.

Since both the reference image and current image are available it is possible to use the efficient second order approximation (ESM) proposed in [11], [3]. In this case the ESM approximation is given as:

$$\boldsymbol{\mathcal{I}}(\tilde{\mathbf{x}}) \approx \frac{1}{2}\left(\boldsymbol{\mathcal{I}}(\mathbf{0}) + \frac{\mathbf{J}(\mathbf{0}) + \mathbf{J}(\tilde{\mathbf{x}})}{2}\tilde{\mathbf{x}}\right)$$

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

The current Jacobian $\mathbf{J}(\mathbf{0})$ is quite straightforward and can be decomposed into modular parts as: $\mathbf{J}(\mathbf{0}) = \mathbf{J}_{\boldsymbol{\mathcal{I}}}\mathbf{J}_{\mathbf{K}}\mathbf{J}_{\mathbf{P}}\mathbf{J}_w\mathbf{J}_{\mathbf{T_c}(\mathbf{0})}\mathbf{J}_{\mathbf{V}}$, where $\mathbf{J}_{\boldsymbol{\mathcal{I}}}$ is the current image-pair gradient of dimension $2n \times 4n$, $\mathbf{J}_{\mathbf{K}}$ is the partial derivative of the pixel coordinates with respect to their corresponding metric coordinates of dimension $4n \times 4n$, $\mathbf{J}_{\mathbf{P}}$ is the partial derivative of each metric coordinates w.r.t the un-normalised point coordinates of dimension $4n \times 6n$, $\mathbf{J}_w$ the Jacobian of the un-normalized point coordinates with respect to the elements warping function of dimension $6n \times 2 * (3 * 3 * 3)$ and $\mathbf{J}_{\mathbf{T_c}(\mathbf{0})}$ is the partial derivative of the tensor elements w.r.t the canonical pose parameters of dimension $2 * 27 \times 2 * 6$. Note that there are two sets of unknown parameters at this stage corresponding to the left and right trifocal tensors respectively.

The last Jacobian $\mathbf{J}_{\mathbf{V}}$ of dimension $2 * 6 \times 6$ is used to center the two components of $\mathbf{J}_{\mathbf{T_c}(\mathbf{0})}$, corresponding to the left and right canonical coordinate systems, so that they represent the same minimal set of unknown parameters. By the definition given in (5) the unknown parameters are defined to be the those which are related to the central pose of the stereo-pair. This Jacobian therefore corresponds to a pair of twist transformation matrices. This twist transformation (the adjoint map), is given as:

$$\mathbf{V} = \begin{bmatrix} \mathbf{R^c} & \mathbf{t^c}_{\times}\mathbf{R^c} \\ \mathbf{0}_3 & \mathbf{R^c} \end{bmatrix}, \quad (10)$$

where $\mathbf{T^c} = (\mathbf{R^c}, \mathbf{t^c})$ is the centering pose given in (4), which maps the current left camera matrix to the stereo center according to (4). Similarly, an adjoint map can be obtained to transform the twist of the right current camera with respect to the right reference camera using $\mathbf{T^{c\prime}}$.

The reference Jacobian $\mathbf{J}(\tilde{\mathbf{x}})$ is obtained as:

$$\mathbf{J}(\tilde{\mathbf{x}}) = \mathbf{J}_{\boldsymbol{\mathcal{I}}^*}\mathbf{J}_{\mathbf{K}}\mathbf{J}_{\mathbf{P}}\mathbf{J}_w\mathbf{J}_{\mathbf{T_c}(\tilde{\mathbf{x}})}\mathbf{J}_{\mathbf{V}},$$

where only $\mathbf{J}_{\boldsymbol{\mathcal{I}}^*}$ and $\mathbf{J}_{\mathbf{T_c}(\tilde{\mathbf{x}})}$ differ from the current Jacobian.

Computing $\mathbf{J}_{\mathbf{T_c}(\tilde{\mathbf{x}})}$ usually requires knowing the solution $\tilde{\mathbf{x}}$ to the estimation problem. However, due to the left invariant structure of the Lie group it can be shown that:

$$\mathbf{J}_{\mathbf{T_c}(\tilde{\mathbf{x}})}\mathbf{J}_{\mathbf{V}}\mathbf{x} = \mathbf{J}_{\mathbf{T_c}(\mathbf{0})}\mathbf{J}_{\mathbf{V}}\mathbf{x}. \quad (11)$$

In this way the ESM second order approximation is given by:

$$\mathcal{J}(\mathbf{x})|_{\mathbf{x}=\tilde{\mathbf{x}}} = \frac{(\mathbf{J}_{\boldsymbol{\mathcal{I}}} + \mathbf{J}_{\boldsymbol{\mathcal{I}}^*})}{2}\mathbf{J}_{\mathbf{K}}\mathbf{J}_{\mathbf{P}}\mathbf{J}_{\mathcal{T}}\mathbf{J}_{T(\mathbf{0})}\mathbf{J}_{\mathbf{V}}, \quad (12)$$

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

The objective function is minimised by iteratively solving (9) by using (12) and (6) for:

$$\mathbf{x} = -\lambda(\mathbf{D}\mathcal{J})^{+}\mathbf{D}(\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{I}}^*), \quad (13)$$

where $(\mathbf{D}\mathcal{J})^{+}$ is the pseudo-inverse, $\mathbf{D}$ is a diagonal weighting matrix determined from a robust function (see [6]) and λ is the gain which ensures an exponential decrease of the error.

A. Reference Image-pairs

As opposed to 3D model-based estimation techniques, no pose initialisation is required and the 3D odometry estimation simply begins at the origin (identity). Dense correspondences are, however, required for each reference image-pair (see Section V-B).

As the camera pair moves through the scene the reference image may be no longer visible or the warped resolution becomes so poor that it is necessary to interpolate many pixels. In both cases this leads to miss-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 reinitialise the tracking. As long as the same reference image is used then the minimisation cut-off thresholds can be tuned for speed since the next estimation will recover any leftover error, however, if the reference image is changed the previous estimate is minimised with smaller cut-off thresholds so as to minimise any drift that may be left over.

V. Implementation

A. 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

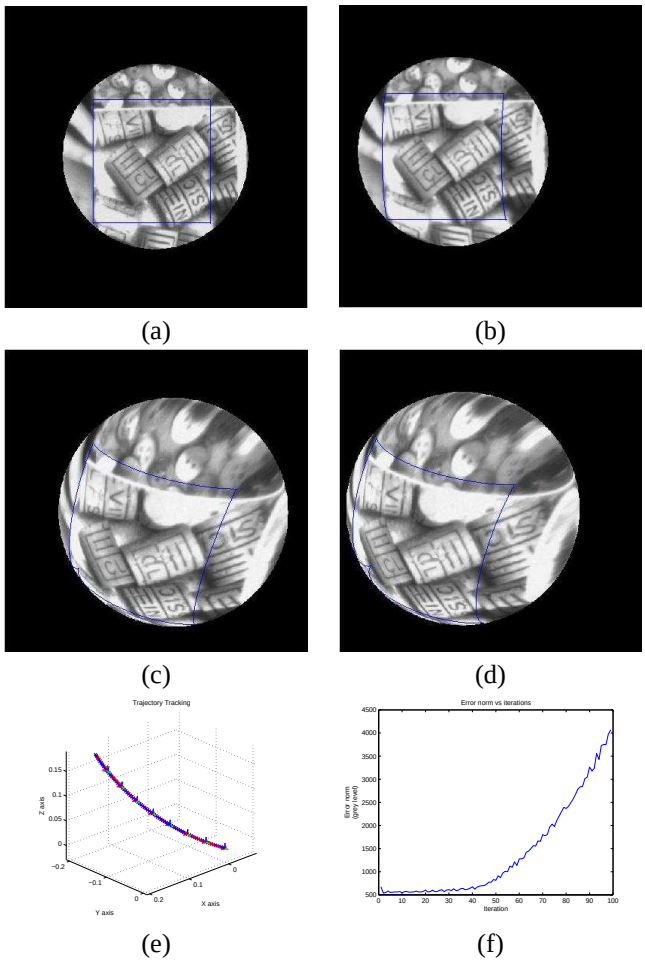


Fig. 2. Simulation of tracking a patch (outlined by the contour) on a 3D sphere: (a,b) Initial left and right images, (c,d) final left and right images, (e) the simulated trajectory of the sphere, (f) the norm of the error in the stereo-image pair.

camera. In order to only test the capability of the proposed tracker the true image correspondences were provided.

In Figure 2 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, these pixels are no longer used for estimation and are rejected by a robust estimator. In (e) one can see the simulated trajectory of the camera pair. In (f) it can be seen that even if the error norm is very small (between 500 and 4000 grey scale levels across the entire stereo-pair) there is an increase in interpolation error as the camera gets further away from the reference image.

B. Dense Correspondences

As mentioned, the reference image pair(s) 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 [15]. 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 [18]). In this paper the approach given in [14] was used, however, any other type of dense correspondence algorithm could be used. The chosen method 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 3 for correspondence results of a typical image pair. In this case the disparity search region was fixed in a range of -20 to -60 pixels along the epipolar lines. Since the baseline is relatively small it is possible to obtain a highly redundant number of dense correspondences.

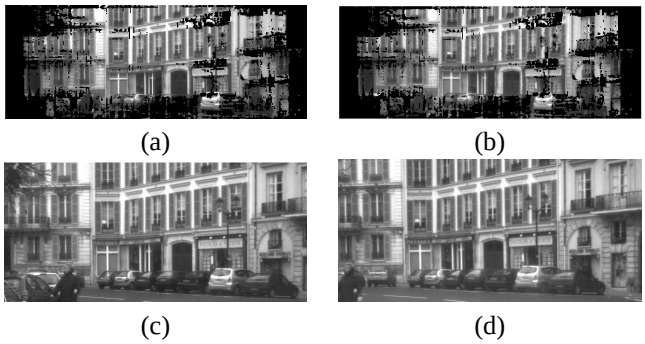
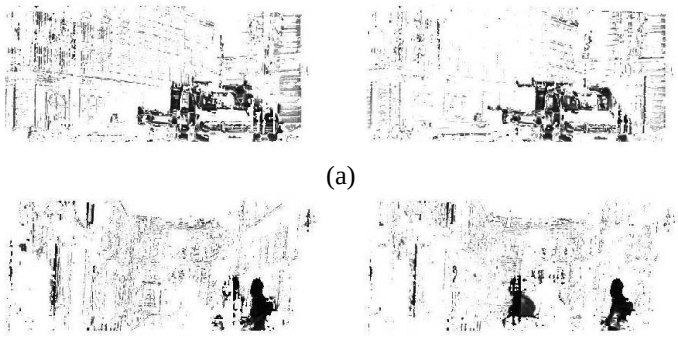


Fig. 3. Dense correspondence of an urban canyon with correspondence occlusion in black: (a) left image and (b) right image warped to the same shape as the left, (c,d) the original images.

C. Robust Estimation

A robust M-estimation technique (as detailed in [5]) was used to reject outliers not corresponding to the definition of the objective function. The use 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 4 (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 has exploited all the useful information in the image so as to estimate the pose. In figure 4 (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 could be useful in an application for determining the trajectory of moving obstacles.



(a)

(b)

Fig. 4. 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.

D. 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 meters per image with a car travelling between 50 and 70 km/hr. Even though tracking succeeds without predictive filtering, in order to improve computational efficiency (significantly less iterations in the minimisation) it was necessary to implement a predictive filter. In this paper the well known Extended Kalman filter described in [19] was used for filtering the pose (i.e. the pose estimate was considered to be the measurement input to the filter).

E. Trajectory Estimation

The algorithm was tested on real full-scale sequences as can be seen in Figure 5 and 6. Radial distortion has been removed from the images before processing. Several test sequences from different streets in Versailles, France, were used to validate the results. These video demonstrations plus more are available online at the authors websites.

The sequence shown in Figure 5 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 $2.9cm/pixel$ for the Versailles region. The path length measured by both Google earth and the tracker was about $440m$. 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 have 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 6, is particularly illustrative since a full loop of the round-about was performed. In particular this enables the drift to be measured at the crossing point in the trajectory. In the case of this round-about the drift at the crossing point was approximately $20cm$ in the vertical direction to the road-plane. Considering that the trajectory around the round-about is approximately $200m$ long (measured using Google earth), this makes a drift of 0.01% on the measurable axis. 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. Due 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.

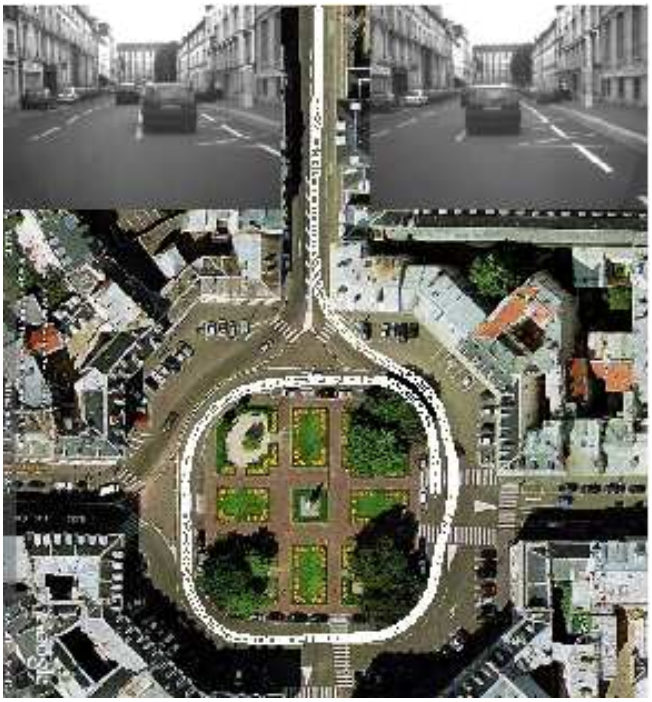
(a)

Fig. 5. Trajectory tracking around along a road in Versailles : (a) The trajectory shown in white has been superimposed on a satellite image. An typical stereo image is shown at the top.

F. Computational requirements

A prototype has been written in Matlab to prove the concept (there is no code or hardware optimization). Even so, it computes on average at approximately 1.5Hz when 10% of the information is used (the strongest gradients) and at around 10sec/image-pair when an image of 759×280 is used. Furthermore, there is only a small difference in precision between the full and reduced images. With the numbers given here there was only about 0.004% drift in translation and $0.03deg/deg$ drift in rotation when measured from a 360m long sequence. The approach is very efficient and could be implemented to run in real-time at video-rate. Two main options are possible:

1. Use an off-line training sequence to obtain a set of dense corresponding reference image-pairs for use online.
2. Perform dense correspondences online (i.e. real-time [18]).



(a)

(b)

(c)

Fig. 6. Trajectory tracking around a round-about 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 round-about (4 points are required to estimate the pose). The length of the path is approximately 392m taken in 698 images. The maximum inter-frame displacement was $1.78m$ and the maximum inter-frame rotation was 2.23^o (b and c) several occlusions which occurred during the sequence and image 300 and 366 respectively (on the right side of the round-about).

VI. CONCLUSIONS AND FUTURE WORKS

The quadrifocal tracking methodology described in this paper has shown to be very efficient, accurate (very small drift) and robust over a wide range of scenarios. The approach is very interesting because trajectory estimation is integrated into a single global sensor-based process that does not depend of intermediate level features. Tracking is initialised automatically at the origin within the visual odometry approach. Furthermore, a compact image-based stereo model of the environment may be obtained using standard dense stereo correspondence algorithms and instead of explicit estimation of an a-priori 3D model. The robust efficient second order minimisation technique also allows minimisation 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 Localisation and Correspondence style approach. It would be interesting to test loop closing procedures and devise strategies to recognise previously seen places within this framework.

REFERENCES

[1] S. Avidan and A. Shashua. Threading fundamental matrices. *Pattern Analysis and Machine Intelligence*, 23(1):73–7, Janvier 2001.

[2] S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. In *Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition*, December 2001.

[3] S. Benhimane and E. Malis. Real-time image-based tracking of planes using efficient second-order minimization. In *IEEE International Conference on Intelligent Robots Systems*, Sendai, Japan, 28 September - 2 October 2004.

[4] A. Chiuso, P. Favaro, H. Jin, and S. Soatto. Structure from motion causally integrated over time. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):523–535, 2002.

[5] A.I. Comport, E. Marchand, and F. Chaumette. Statistically robust 2d visual servoing. *IEEE Transactions on Robotics*, 22(2):415–421, April 2006.

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

[7] A. J. Davison and D. W. Murray. Simultaneous localisation and map-building using active vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 2002.

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

[9] R. Hartley and A. Zisserman. *Multiple View Geometry in computer vision*. Cambridge University Press, 2001. Book.

[10] P.-J. Huber. *Robust Statistics*. Wiler, New York, 1981.

[11] E. Malis. Improving vision-based control using efficient second-order minimization techniques. In *IEEE International Conference on Robotics and Automation, ICRA'04*, volume 2, pages 1843– 1848, New Orleans, April 26-May 1 2004.

[12] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Real-time localization and 3d reconstruction. In *IEEE Conference of Vision and Pattern Recognition*, New-York, USA, June 2006.

[13] D. Nistér, O. Naroditsky, and J. Bergen. Visual odometry. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 652–659, CVPR 2004, July 2004.

[14] A.S. Ogale and Y. Aloimonos. Shape and the stereo correspondence problem. *International Journal of Computer Vision*, 65(1), October 2005.

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

[16] A. Shashua and L. Wolf. On the structure and properties of the quadrifocal tensor. In *European Conference on Computer Vision*, pages 710–724, 2000.

[17] N. Simond and P. Rives. Trajectography of an uncalibrated stereo rig in urban environments. In *IEEE RSJ/International conference on Intelligent Robot and System, IROS*, pages 3381–3386, Sendai, Japan, 28 September - 2 October 2004.

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

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