Real-time Dense Visual Tracking under Large Lighting Variations

Maxime Meilland, Andrew I. Comport, Patrick Rives

Real-time Dense Visual Tracking under Large Lighting Variations

Maxime Meilland[1]
maxime.meilland@inria.fr

Andrew Ian Comport[2]
comport@i3s.unice.fr

Patrick Rives[1]
patrick.rives@inria.fr

[1] INRIA Sophia Antipolis Méditerranée
2004 Route des Lucioles BP 93
Sophia Antipolis, France

[2] CNRS-I3S, UNSA
2000 Route des Lucioles BP 121
Sophia Antipolis, France

Abstract

This paper proposes a model for large illumination variations to improve *direct* 3D tracking techniques since they are highly prone to illumination changes. Within this context dense monocular and multi-camera tracking techniques are presented which each perform in real-time (45Hz). The proposed approach exploits the relative advantages of both model-based and visual odometry techniques for tracking. In the case of direct *model-based* tracking, photometric models are usually acquired under significantly greater lighting differences than those observed by the current camera view, however, model-based approaches avoid drift. Incremental *visual odometry*, on the other hand, has relatively less lighting variation but integrates drift. To solve this problem a hybrid approach is proposed to simultaneously minimise drift via a 3D model whilst using locally consistent illumination to correct large photometric differences. Direct 6 dof tracking is performed by an accurate method, which directly minimizes dense image measurements iteratively, using non-linear optimisation. A stereo technique for automatically acquiring the 3D photometric model has also been optimised for the purpose of this paper. Real experiments are shown on complex 3D scenes for a hand-held camera undergoing fast 3D movement and various illumination changes including daylight, artificial-lights, significant shadows, non-Lambertian reflections, occlusions and saturations.

1 Introduction

This work is supported through a project for automatic visual navigation in structured complex 3D environments, however, as will be shown, the application domain is much wider and also includes robot localisation and augmented reality. This paper focuses on the real-time tracking part of visual navigation in the presence of significant illumination change using an appearance-based method. *Direct* tracking methods [2] use an estimation model that minimises an error directly in the sensor space. This allows them to be robust to modeling error and to provide a single simplified model that does not require propagating uncertainty between different stages (such as feature based approaches). Whilst direct tracking approaches are incremental and local, full-image stereo techniques have been shown to efficiently track

BMVC 2011 http://dx.doi.org/10.5244/C.25.45

over a large convergence domain, especially when coupled with a multi-resolution pyramid [6]. The main drawback and criticism with this type of approach is therefore that of being able to handle significant illumination variation.

To handle illumination for direct methods, several approaches have been proposed. In [3] an affine model is used but only global changes are modeled. In [16] a more complete model is employed, including diffuse and specular reflections. Very good results are obtained, but a lot of unknowns are introduced in the minimization process which makes the algorithm more complex and robustness to occlusions is not handled. In [9] a complete robust registration technique is presented. Illumination changes are handled by modeling image variation using low dimensional linear subspaces. A learning step is carried out to generate invariant image representations from a set of images taken under different illuminations. Despite this learning step, the model construction deals only with Lambertian surfaces and is sensitive to occlusions and self-shadowing.

In the multi-view reconstruction and SFM literature [15], powerful models have been used to perform 3D reconstruction. Illumination is generally modeled using surface normals and albedo [4, 17] but most of that work only considers Lambertian surfaces and distant lights, which are clearly not valid in uncontrolled environments. Furthermore those methods are often complex and used offline and are not well suited for online camera tracking.

Instead of using *direct* Sum of Squared Differences (SSD) [2], other error functions have been used. In [12] normalized cross correlation (NCC) is employed jointly with gradient oriented pyramids. This allows to register different images, however, NCC is a local measure and is not well suited for 3D scenes where perspective effects are significant. Other authors have proposed to maximize the information shared between images using Mutual Information [7, 14]. This is currently, one of the best ways to register highly different multimodal images (i.e. between a map and a satellite image). Nevertheless, these methods rely on smoothing an uncontinuous cost function in order to obtain an analytic gradient which degrades precision and leads to uncertain performance and convergence properties. Furthermore these methods rely on maximizing an information metric which is not sensor-based (and therefore they are less robust to model error).

To the authors' knowledge, the existing real-time *direct* tracking techniques which consider light change only tackle simple geometric scene models: planar homographies [9, 16] or cylinders [5] and only small regions (patches) are tracked. Alternatively, the method proposed here handles general 3D scenes, and many additional undesirable effects are present: occlusions, self-shadowing, objects interactions, specular reflections, camera saturation and also uncertainties on the scene geometry (*e.g.* due to reconstruction). The central purpose of this paper is to address these illumination problems whilst maintaining a *dense* direct approach and considering real-time aspects.

1.1 Proposed approach

The novelty of this paper is to propose a hybrid model-based/visual-odometry method which is robust to a large set of illumination changes whilst minimising drift with respect to a global reference frame (3D model). Firstly, a model-based (MB) approach is considered here to be one that minimises the error between a known model (3D+photometric) and the warped current image. In terms of illumination, large changes can be expected for the MB approach (see Fig. 1(b)) since there is a large temporal difference between the moment the model was acquired and the current image. Secondly, a visual odometry (VO) approach is defined as minimising the error between the image acquired at time $t-1$ and the warped current image at time t (see Fig. 1(a)). In this case relatively small changes can be expected (even if they must still be modeled). The following four types of illumination changes are identified:

- **Global-Step**. Abrupt lighting changes across the entire image - an ambient light being switched on or a large difference when initialising the pose between a model and a camera.
- **Local-Step**. Abrupt lighting changes locally in the image - a spotlight being switched on, specular reflections, shadows or occlusions induced by a change of camera viewpoint.
- **Global-Ramp**. Slow lighting changes across the entire image - a moving light or a change of light source intensities (*e.g*. clouds, position of the sun,...).
- **Local-Ramp**. Slow lighting changes locally in the image - a moving spotlight, shadows or local reflections.

The MB approach can be considered to be in a permanent *Step* configuration for local and global variation (although some local gradient must be preserved). Nevertheless, it is this approach which allows to avoid drift. The VO technique, on the other hand, undergoes more incremental lighting change (*Ramp* + *Step*) and is therefore much more robust to changes but it accumulates drift over time. As in [8], robustness to *Global* variation is handled by a robust centering measure and *Local-step* variations are rejected along with outliers using robust M-estimation. This paper, however, only dealt with planar VO tracking.

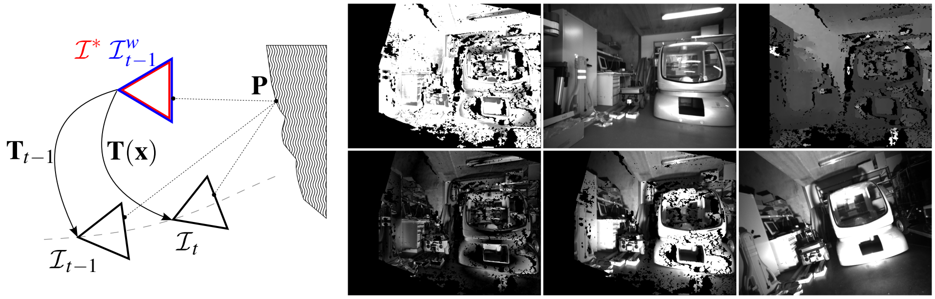
Moreover, a full six degrees of freedom tracking approach is proposed which is based on a 3D model of the environment that contains both photometric and geometric information. The 3D model is automatically reconstructed in near real-time from a learning sequence that has been acquired by a stereo acquisition system. Online real-time tracking is then performed at 45Hz using a monocular or multi-camera system. The proposed approach will be shown to work in real-time with many different illumination conditions and with no drift.

2 Robust 3D direct image registration

Let $\mathcal{I}$ denote a set (considered also as a vector) of brightness image measurements, where for each image pixel $\mathbf{p} \in \mathcal{I}$ a corresponding 3D point $\mathbf{P} \in \mathbb{R}^3$ is available (*e.g*. extracted offline by dense stereo matching). It is assumed that the scene geometry does not vary in time. The set of the measurements $\mathbf{S} = \{\mathcal{I}^*, \mathcal{P}^*, \mathbf{W}\}$, denoted with a '$*$', will be referred throughout that paper as the *augmented reference image*, where $\mathcal{P}^* = \{\mathbf{P}_1, \ldots, \mathbf{P}_n\}$, n is the number of pixels in the reference image and $\mathbf{W} \in \mathbb{R}^{1\times n}$ is a saliency map indicating which pixels condition each of the 6 dof best.

Multiple augmented reference images are then joined together to form a global 3D photometric model. The N augmented reference images are related via a graph $\mathcal{G} = \{\mathbf{S}_1, \ldots, \mathbf{S}_N, \mathbf{T}_1, \ldots, \mathbf{T}_M\}$, of M poses, which have been estimated off-line using the approach in [6] which will be re-introduced in Section 2.2. It can be noted that very few works have appeared in the literature which perform direct dense tracking from arbitrary geometry due to the difficulty in acquiring such models. Here this stereo technique has been made to run in near real time for fast augmented model acquisition (see Sub-section 3.2).

In the following subsections, first a model-based tracking approach will be presented and then a visual odometry based approach will be presented before combining both within a hybrid tracking framework.



(a) Cameras position in the scene.

(b) MB tracking: registration results.

Figure 1: (a). In red, the augmented reference image $\mathcal{I}^*$. In blue, the image $\mathcal{I}^w_{t-1}$ from time $t-1$ warped onto the reference. The current camera image $\mathcal{I}_t$ at time t. For the MB configuration the error $\mathcal{I}_t - \mathcal{I}^*$ is minimised and for the VO configuration $\mathcal{I}_t - \mathcal{I}^w_{t-1}$ is minimised. (b). *Top left*, robust outliers weights. *Top center*, augmented reference image $\mathcal{I}^*$. *Top right*, reference depthmap. *Bottom left*, intensity error $\mathcal{I}^w_t - \mathcal{I}^*$ after alignment. *Bottom center*, warped current image $\mathcal{I}^w_t$. *Bottom right*, original current image $\mathcal{I}_t$.

2.1 Model-based (MB) tracking under illumination variation

Assuming that an augmented reference model $\mathcal{G}$ has already been reconstructed, the aim of model-based tracking estimate the 6dof. *motion increment* $\mathbf{x} \in \mathbb{R}^6$ which associates the current image $\mathcal{I}_t$ acquired at time t, with the reference image $\mathcal{I}^*$. $\mathbf{x}$ is related to the homogeneous camera pose matrix $\mathbf{T}$ defined as:

$$\mathbf{T}(\mathbf{x}) = e^{[\mathbf{x}]_\wedge} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \in \mathbb{SE}(3), \quad \mathbf{x} = \int_0^1 (\boldsymbol{\omega}, \boldsymbol{\upsilon})dt \in \mathfrak{se}(3), \qquad [\mathbf{x}]_\wedge = \begin{bmatrix} [\boldsymbol{\omega}]_\times & \boldsymbol{\upsilon} \\ \mathbf{0} & 0 \end{bmatrix}, \tag{1}$$

where $\boldsymbol{\omega}$ and $\boldsymbol{\upsilon}$ are the angular and linear motion components which are integrated to obtain a pose $\mathbf{T}$ and where $[.]_\times$ represents the skew symmetric matrix operator, $\mathbf{R} \in \mathbb{SO}(3)$ is a rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ is a translation vector.

The warping function $\mathbf{p} = w(\mathbf{P}, \mathbf{K}; \mathbf{T}(\mathbf{x})) = \mathbf{K}[\mathbf{R} \;\; \mathbf{t}]\mathbf{P}$ transfers 3D Euclidean points $\mathbf{P}$ onto another camera plane, via perspective projection according to a camera motion $\mathbf{x}$, where $\mathbf{K} \in \mathbb{R}^{3\times 3}$ is the upper triangular intrinsic parameter matrix of the current sensor, which is calibrated and where $\mathbf{p}$ is the homogeneous pixel vector.

In the case of the true motion $\bar{\mathbf{x}}$, the current image intensities $\mathcal{I}_t$, at time t, are related to the reference image $\mathcal{I}^*$ by:

$$\mathcal{I}^*(\mathcal{P}^*) = \boldsymbol{\alpha}_{MB}\Big(\mathcal{I}_t\big(w\big(\mathcal{P}^*; \mathbf{T}(\bar{\mathbf{x}})\big)\big)\Big) - \beta_{MB}, \quad \forall \mathcal{P}^* \in \mathbb{R}^3, \tag{2}$$

where the corresponding intensities are interpolated bi-linearly at points $\mathbf{p}$ to create novel view [1]. $\boldsymbol{\alpha}_{MB} = diag(\alpha_1, \ldots, \alpha_i) \in \mathbb{R}^{n\times n}$ is the local affine pixel-by-pixel diagonal gain matrix and $\beta_{MB} \in \mathbb{R}$ is the global bias shift in intensity due to global illumination changes.

The current camera pose can be estimated by minimizing the intensity error between the warped current image and the reference.

$$\mathbf{e}_{MB} = \rho\left(\boldsymbol{\alpha}_{MB}\mathcal{I}_t\left(w(\mathcal{P}^*; \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}))\right) - \beta_{MB} - \mathcal{I}^*(\mathcal{P}^*)\right), \tag{3}$$

where $\widehat{\mathbf{T}}$ is the estimated pose up to time $t-1$ and $\mathbf{T}(\mathbf{x})$ is the incremental pose at time t.

Since this is a non-linear function in the unknown pose parameters, a robust iteratively re-weighted least squares minimization procedure can be used [6] to minimise the error function (3) $\mathcal{O}(\mathbf{x}) = \underset{\mathbf{x}}{\arg\min}\,(\mathbf{e}_{.}(\mathbf{x}))$. An inverse compositional algorithm is used [2], which allows to precompute most of the minimization parts directly in the augmented reference image. In this case the unknown $\mathbf{x}$ is iteratively updated using Gauss Newton optimisation:

$$\mathbf{x} = -(\mathbf{J}_{MB}^T \mathbf{D}_{MB} \mathbf{J}_{MB})^{-1} \mathbf{J}_{MB}^T \mathbf{D}_{MB} \mathbf{e}_{MB}, \tag{4}$$

where $\mathbf{J}_{MB}$ is the Jacobian of (3) wrt. $\mathbf{x}$.

As mentioned in the introduction, the MB approach is considered to be continually in a global step configuration since it was acquired under significantly different illumination conditions than the current camera. The *global* illumination change is determined robustly by performing a global shift of the error with respect to a *robust* estimate of the center of the distribution (its median): $\beta_{MB} = Median(\mathbf{e}_{MB})$. This efficient and robust technique was originally proposed in [8] which the interested reader may consult for more detail.

On the other hand, the pixel-by-pixel gain $\boldsymbol{\alpha}_{MB}$ (*Local* illumination variation) is, unfortunately, not locally observable and it is difficult to separate from local occlusions and interpolation error. For the purpose of model-based tracking, the local gain $\boldsymbol{\alpha}_{MB}$ will be considered to be absorbed by a robust diagonal weighting matrix $\boldsymbol{\alpha}_{MB} \subset \mathbf{D}_{MB} = diag(w_1, \ldots, w_i)$. The weights w_i, which reflect the confidence of each pixel, are given by [11]:

$$w_i = \frac{\psi(\delta_i/\sigma)}{\delta_i/\sigma}, \quad \psi(u) = \begin{cases} u & , \text{ if } |u| \le a \\ a\frac{u}{|u|} & , \text{ if } |u| > a, \end{cases} \tag{5}$$

where δ_i is the normalized residue given by $\delta_i = \Delta_i - Med(\Delta)$ and ψ is the influence function.Huber's function has been chosen where the proportionality factor is $a = 1.2107$ and represents 95% efficiency in the case of Gaussian noise.

It will be seen in the results that since the MB (3) images differ greatly from the current image, convergence is slow, there is a greater potential for local minima and tracking is not robust and easily fails.

2.2 Visual odometry(VO) tracking under illumination variation

This section describes a non-classic visual odometry function that takes advantage of the precomputed depth map from a known 3D model $\mathcal{G}$. This allows to avoid computing a dense depth map online which is very costly for real-time applications. The modified VO technique is therefore based on the augmented reference image $\mathbf{S} = \{\mathcal{I}_{t-1}^w, \mathcal{P}^*, \mathbf{W}\}$ which combines the 3D model reference points $\mathcal{P}^*$ with the image $\mathcal{I}_{t-1}^w$ from time $t-1$. This image is warped, in conjunction with the current image, to the reference image position (see Fig. 1(a)):

$$\mathcal{I}_{t-1}^w(\mathcal{P}^*) = \mathcal{I}_{t-1}(w(\mathcal{P}^*; \mathbf{T}_{t-1})) \tag{6}$$

This leads to an optimised VO error function (corresponding to the temporal inter-frame intensity variation) that is expressed in the reference frame:

$$\mathbf{e}_{VO} = \rho\left(\boldsymbol{\alpha}_{VO}\mathcal{I}_t(w(\mathcal{P}^*; \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}))) - \beta_{VO} - \mathcal{I}_{t-1}^w(\mathcal{P}^*)\right), \tag{7}$$

where $\mathbf{T}_{t-1}^{-1}\mathbf{T}(\mathbf{x})$ is the classic inter-frame VO camera motion. The illumination changes have been modeled again by a global bias β_{VO} and local gain $\boldsymbol{\alpha}_{VO} \subset \mathbf{D}_{VO}$. It is assumed that the pose of the camera $\mathbf{T}_0$ at time $t = 0$ has been adequately initialised (either by an initialisation procedure or initial conditions that are within the convergence domain). (7) is minimized using a robust least square optimisation as in (4).

Considering that images are acquired at video framerate (*e.g.* ≥ 25 Hz), the illumination changes for VO between two consecutive frames $\mathcal{I}_{t-1}$ and $\mathcal{I}_t$ are much smaller than for the MB technique (even if some step variation is allowed) : $\boldsymbol{\alpha}_{VO} << \boldsymbol{\alpha}_{MB}$ and $\beta_{VO} << \beta_{MB}$ meaning that the error $\mathbf{e}_{VO} << \mathbf{e}_{MB}$. Since the global step change is easily modeled, both VO and MB approaches handle nicely global variation. On the other hand, the temporally local illumination change can be more efficiently handled by the VO approach and in many cases the MB approach will simply fail due to heavy local variation (see Fig. 3). Once again this approach still allows to account for outliers.

2.3 Hybrid model-based and visual odometry (H) tracking

This section presents a hybrid approach to take advantage of both MB and VO approaches. In the first case MB tracking ensures no drift, however, it is prone to large illumination differences and subsequently slow to failed convergence. On the other hand, VO provides temporally close illumination measurements, however, it is prone to drift. In a sensor-based approach it is important to maintain raw sensor measurement as reference in the minimisation process to ensure precision, avoid drift and remain robust to modeling error. In that case, if an illumination model was to incrementally adjust the reference image then drift would creep into the system. It is for this reason that the local illumination changes are not estimated as part of the state vector.

In this hybrid case a global error function is defined by stacking $\mathbf{e}_{MB}$ and $\mathbf{e}_{VO}$ as:

$$\mathbf{e}_H = \begin{bmatrix} \rho\left(\mathcal{I}_t(w(\mathcal{P}^*;\mathbf{T}(\mathbf{x})) - \mathcal{I}^*(\mathcal{P}^*) - \beta_{MB}\right) \\ \rho\left(\mathcal{I}_t(w(\mathcal{P}^*;\mathbf{T}(\mathbf{x})) - \mathcal{I}^w_{t-1}(\mathcal{P}^*) - \beta_{VO}\right) \end{bmatrix}. \tag{8}$$

Since the minimized error distributions $\mathbf{e}_{MB}$ and $\mathbf{e}_{VO}$ may have different modalities, two independent robust and centering functions are used, to ensure not rejecting the non-dominant modality. (8) is minimized using a robust least square optimisation as in (4). The hybrid Jacobian is efficient to compute since the geometric part of the VO is the same as for MB.

Fig. 2 shows an example convergence obtained with MB-VO minimization (8) compared to MB (3) and VO (7). Since the VO inter-frame error is close to zero, a fast exponential convergence to the minimum is guaranteed, but small errors are accumulated. The MB approach convergence is slower due to large illumination differences whilst the hybrid method (H) converges faster and without drift.

3 Results

3.1 Experimental validation

To prove the efficiency and the robustness of the method, a set of augmented reference images are acquired and reconstructed from a stereo image pair (see Section 3.2). The current image sequence was acquired using a hand held camera and extreme light variations are created in several ways including using a spotlight to generate local step and ramp variations, whilst creating shadows due to occluding objects and sensor saturations in the current image.

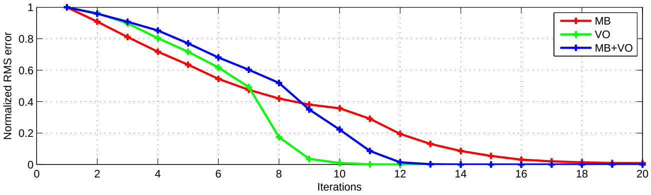


Figure 2: Convergence speed is compared between MB (red), VO (green) and the proposed hybrid H (blue).

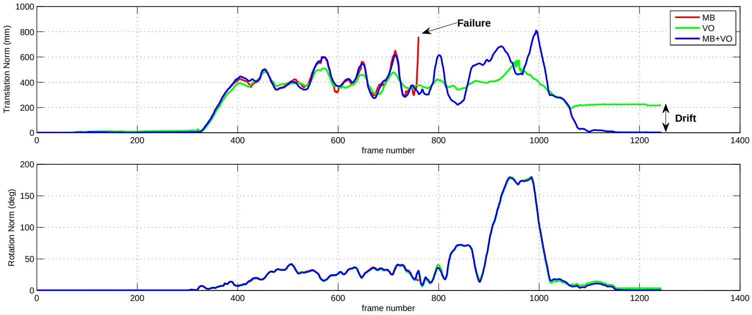


Figure 3: Norms of the 6 estimated dof of a loop trajectory containing 1243 images. *Top*, translation norm. *Bottom*, rotation norm. The MB tracking fails after image 761. Error was accumulated using only VO. The proposed hybrid approach (H) tracked well the camera, giving an identity pose at the end of the sequence.

At time $t = 0$ the current image pose w.r.t the reference pose is initialised at a known starting point so that $\mathbf{T}(\mathbf{x}) = \mathbf{I}$. Movements of the camera are generated manually thought the full six dof and the camera is returned to the origin (a camera tripod) so as to provide a ground truth on the final camera pose (i.e. the identity matrix).

One of the test sequences presented here contains a set of 1243 images. The sequence was tracked using 3 different methods. First only the MB (3) approach was performed, next only the VO (7) approach and finally using the proposed hybrid approach (H) (8). Each tracking phase was computed using the same tuning parameters (*e.g.* maximum number of iterations, convergence threshold).

Fig. 3 shows the six estimated dof obtained using each method. It can be seen that for the MB method, the tracking fails at image 761. The VO method tracked throughout the image sequence but errors are accumulated along the sequence leading to drift. The hybrid method, tracked well throughout the entire image sequence. No drift is accumulated since the true reference image is still involved non-linearly in the minimization process.

3.2 Automatic 3D Model acquisition

To build the 3D model $\mathcal{G}$ it is necessary to perform dense localisation and mapping during a learning phase. In practice a classic rectified stereo rig is used. Dense matching is then computed along epipolar scan-lines using a standard algorithm [10] and triangulation is performed to obtain a set of 3D points $\mathcal{P}^*$ corresponding to the image pixels.

Since the 3D model is reconstructed offline it is possible to use a more accurate and robust parametrisation of the localisation method. In particular, dense optimisation is performed using all the available data, estimation is allowed more iterations to converge and loop closing is performed where possible. Direct minimization is performed between the last constructed augmented reference image and the current image in the sequence. A new reference image is constructed and added to the model when the robust estimate of the scale of the minimized error distribution (*e.g.* the median absolute deviation (MAD)), is greater than a certain threshold or when the number of overlapping pixels between the current and the last reference view is too small.

3.3 Real-time Experiments

In each experiment, a 3D model is first acquired as in Section 3.2. In the experiment presented here, the 3D model is built and 20 reference images were necessary to cover the zone which was mapped. This database is later used to track a hand-held monocular camera (or stereo camera pair) in real-time. Using stereo online is more robust, however, it requires slightly more computation than monocular tracking.

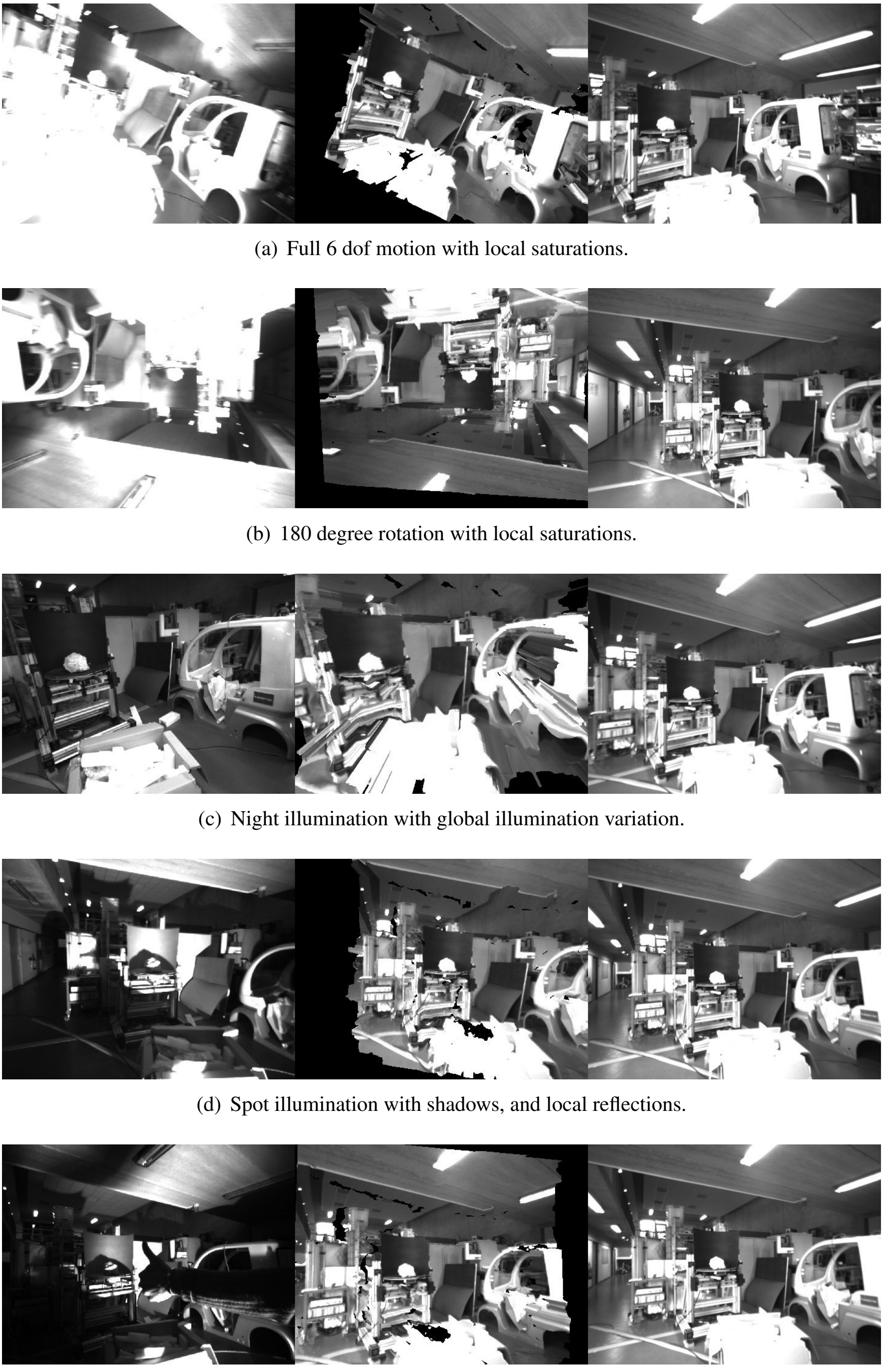
For the *online* tracking, a real-time implementation has been realized in C++. The algorithm runs at 45 Hz on an Intel Core 2 Duo laptop at (2.2 Gz) for stereo images of dimension 800×600. To achieve this frame rate optimisation was first performed by creating a saliency map of the reference images' pixels [13]. The saliency map was then used to select a small number of the pixels necessary to accurately perform the tracking (about 3.10^4 pixels/image).

To initialise tracking (or recover from tracking failure), the closest image from the database is obtained by performing direct low-resolution tracking between the first image acquired from the live camera and each image from the database. The best score in terms of "Zero Normalised Correlation" is then chosen as the initial image. The initial pose $\mathbf{T}_0$ is then estimated at full resolution. To guarantee convergence and precision at initialisation, all pixels are used in the minimization and a lot of iterations are allowed. Since this is an initialization step, more computation time is allowed and the camera is held static for a few seconds. A technique which scales well with large models would merit further investigation.

The main set of results captured from the real-time tracking experiments are shown in Fig. 4. The figure shows some aligned images from different tracking experiments. In the figure a first experiment was performed in the afternoon ((a), (b)) with direct beams of sunlight modifying illumination conditions. Another experiment was performed at night ((c), (d), (e)) and only artificial lighting was used. The camera exposure time was intentionally set high in order to provoke camera saturations. To highlight tracking results, a synthetic view of the reference image is rendered in real-time according to the current computed pose and displayed beside the current image. It can be seen that the proposed tracker is able to track all the 6 dof of the camera in various extreme perturbations cases: camera shaking and motion blur, defocus and aperture changes, occlusions, saturation, shadows and even with 180 degrees rotation. The video accompanying this paper better illustrates the results and is directly captured from the visual output of the tracking implementation, as are the images of Fig. 4).

4 Conclusions

The real-time hybrid tracking method presented here is shown to track robustly in complex 3D scenes at 45Hz on very challenging image sequences. It efficiently combines a model-



(a) Full 6 dof motion with local saturations.

(b) 180 degree rotation with local saturations.

(c) Night illumination with global illumination variation.

(d) Spot illumination with shadows, and local reflections.

(e) Spot illumination with occlusions.

Figure 4: Different tracking results. For each row of the image from left to right: Current image; Synthetic 3D view rendered at the estimated pose (corresponding to the current image pose); Original reference image

based approach, based on a database of augmented images, with an online visual odometry technique. The model-based approach allows to avoid drift while the visual odometry ensures accurate tracking of a camera undergoing large local and global illumination changes. Since this model is non-linear and direct it also leads to accurate tracking. The 3D model was also acquired automatically in near real-time using a stereo learning phase. Salient pixel selection has allowed this direct algorithm to run at high frequency on a standard laptop.

An important aspect which has not been considered in this paper are the geometric changes in the scene over time. A possible future direction will be to extract depth online from a stereo camera pair and jointly minimizing the inter-frame displacement with an augmented reference image from the database.

References

[1] S. Avidan and A. Shashua. Novel view synthesis in tensor space. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1034, 1997. ISSN 1063-6919.

[2] S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 1090, December 2001.

[3] A. Bartoli. Groupwise geometric and photometric direct image registration. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(12):2098 –2108, 2008. ISSN 0162-8828.

[4] Ronen Basri and David Jacobs. Photometric stereo with general, unknown lighting. In *In IEEE Conference on Computer Vision and Pattern Recognition*, pages 374–381, 2001.

[5] Marco La Cascia and Stan Sclaroff. Fast, reliable head tracking under varying illumination. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:322–336, 1999.

[6] A.I. Comport, E. Malis, and P. Rives. Real-time quadrifocal visual odometry. *In The International Journal of Robotics Research*, 29(2-3):245–266, February 2010.

[7] A. Dame and E. Marchand. Accurate real-time tracking using mutual information. In *IEEE Int. Symp. on Mixed and Augmented Reality, ISMAR'10*, pages 47–56, Seoul, Korea, October 2010.

[8] T. Gonçalves and A.I Comport. Real-time direct tracking of color images in the presence of illumination variation. In *IEEE International Conference on Robotics and Automation*, May 2011.

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

[10] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:328–341, 2008.

[11] P.J. Huber. *Robust Statistics*. New york, Wiley, 1981.

[12] M. Irani and P. Anandan. Robust multi-sensor image alignment. In *Computer Vision, 1998. Sixth International Conference on*, pages 959 –966, January 1998.

[13] M. Meilland, A.I. Comport, and P. Rives. A spherical robot-centered representation for urban navigation. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 5196 –5201, 2010.

[14] Giorgio Panin and Alois Knoll. Mutual information-based 3d object tracking. *International Journal of Computer Vision*, 78:107–118, 2008.

[15] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms, 2006.

[16] G. Silveira and E. Malis. Real-time visual tracking under arbitrary illumination changes. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –6, 2007.

[17] Li Zhang, Brian Curless, Aaron Hertzmann, and Steven M. Seitz. Shape and motion under varying illumination: Unifying structure from motion, photometric stereo, and multi-view stereo. In *The 9th IEEE International Conference on Computer Vision*, pages 618–625, Oct. 2003.

[1]

[1] This work has been supported by ANR (French National Agency) CityVIP project under grant ANR-07_TSFA-013-01.