An asymmetric real-time dense visual localisation and mapping system

Andrew I. Comport
CNRS-I3S, UNSA
2000 Route des Lucioles BP 121
Sophia Antipolis, France
comport@i3s.unice.fr

Maxime Meilland and Patrick Rives
INRIA Sophia Antipolis Méditerranée
2004 Route des Lucioles BP 93
Sophia Antipolis, France
firstname.lastname@inria.fr

Abstract

This paper describes a dense tracking system (both monocular and multi-camera) which each perform in real-time (45Hz). The proposed approach combines a prior dense photometric model with online visual odometry which enables handling dynamic changes in the scene. In particular it will be shown how the technique takes into account large illumination variations and subsequently improves direct *tracking techniques which are highly prone to illumination change. This is achieved by exploiting 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]

1. Problem Statement

This demonstration will show 45Hz camera tracking w.r.t a global reference 3D model using a dense *direct* method which minimises directly intensity errors between images [1, 2, 6]. Several performance criteria will be addressed:

- A dense method which is accurate and robust to model uncertainty.
- A method able to cover large scale environments $>$1km.
- Can handle complex geometry of the scene, uncertainties and occlusions.
- Non-linear saliency maps are determined for choosing the pixels which best condition the estimation model.
- *Real-time* tracking needs an efficient illumination model able to handle: Diffuse daylight changes, specular reflections, saturations, self-shadowing ...

2. Direct 3D Model-based (MB) tracking

The unknown 3D motion (translation and rotation) $\mathbf{x}$ between an augmented reference image $\boldsymbol{S} = \{\boldsymbol{\mathcal{I}}^*, \boldsymbol{\mathcal{P}}^*\}$ and the current camera $\boldsymbol{\mathcal{I}}_t$ can be iteratively estimated by minimising a robust error between the warped image and the reference image:

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

- *Global* illumination change is *efficiently* determined using robust metric by performing a global shift of the error: $\beta_{MB} = Median(\mathbf{e}_{MB})$ [3].
- *Local* illumination changes are absorbed by the robust diagonal weighting matrix $\mathbf{D}$ [4].

Inconveniences: Convergence is slow when the MB images differ greatly from the current image and tracking can fail in real-time conditions.

3. Visual odometry(VO) tracking

To improve convergence speed and robustness to local dynamic changes, a non classic visual odometry approach

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

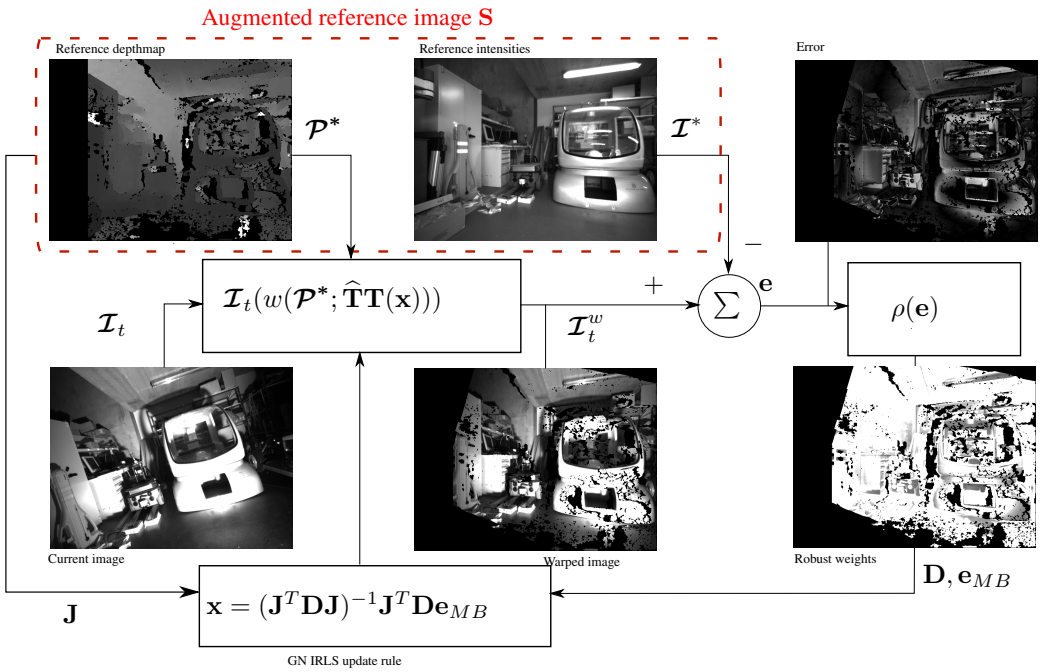


Figure 1. This figure shows the direct non-linear iteratively re-weighted least squared estimation of the pose *Top left*, Reference depth map determined by dense matching, *Top middle*, The reference image $\mathcal{I}^*$, *Top right*, Intensity error $\mathcal{I}_t^w - \mathcal{I}^*$ after alignment, *Bottom left*, Current image $\mathcal{I}_t$, *Bottom middle*, Warped current image $\mathcal{I}_t^w$, *Bottom right*, robust outliers weights.

is proposed. This technique minimises the inter-frame temporal intensities using the geometry of the model and the previous image intensities:

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

Advantages

- 3D geometry is never recomputed since it is shared with the original model.
- Very small local illumination changes can be expected between successive frames (ie. $\geq$ 20Hz) $\Rightarrow$ fast convergence.
- Still robust to global illumination changes (due to the global bias model).

however, **visual odometry *drifts*** with time.

4. Hybrid(H) tracking

The proposed method depicted in Figure 4 performs both model-based and visual odometry tracking simultaneously. Global minimisation of the error functions therefore combines the advantages of each technique.

$$\mathbf{e}_H = \begin{bmatrix}\mathbf{e}_{MB} & \mathbf{e}_{VO}\end{bmatrix}^T.$$

- **Fast** convergence (due to VO).
- **No drift** since raw sensor measurement is maintained in the minimisation process (due to MB).

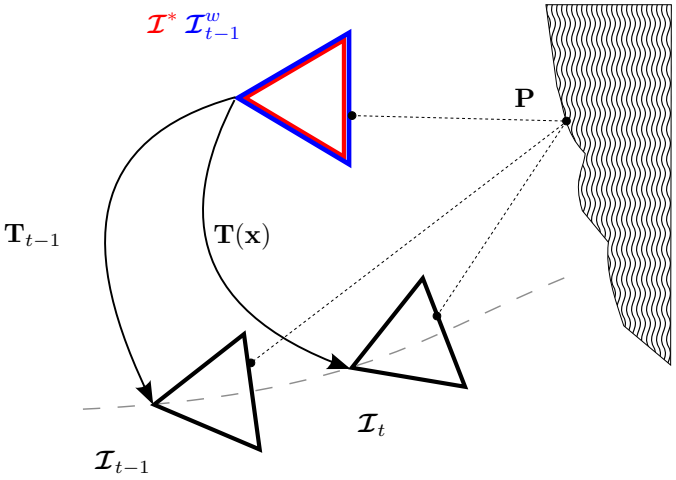


Figure 2. In red, the augmented reference image $\mathcal{I}^*$. In blue, the image $\mathcal{I}_{t-1}^w$ from time $t-1$ warped onto the reference. The current camera image $\mathcal{I}_t$ at time t. For the MB configuration the error $\mathcal{I}_t - \mathcal{I}^*$ is minimised and for the VO configuration $\mathcal{I}_t - \mathcal{I}_{t-1}^w$ is minimised.

5. Robust Non-linear minimisation

Unknown motion $\mathbf{x}$ estimated by an iterative re-weighted least square minimisation:

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

$$w_i = \frac{\psi(\delta_i/\sigma)}{\delta_i/\sigma}, \quad \psi(u) = \begin{cases} u & , \text{ if } |u| \leq a \\ a\frac{u}{|u|} & , \text{ if } |u| > a, \end{cases} \quad (2)$$

6. Scene map

The reconstructed scene map contains the following elements which are depicted in Figure 3:

- $\mathcal{I}^*$: Reference image intensities.
- $\mathbf{P} \in \mathcal{P}^*$: Reference 3D point.
- $\mathcal{I}_t$: Current image at time t.
- $\mathcal{I}_{t-1}$: Last image at time $t-1$.
- $\mathcal{I}_{t-1}^w$: Last warp image at time $t-1$.
- $\mathbf{T}_{t-1}$: Last estimated pose.
- $\mathbf{T}(\mathbf{x})$: Unknown pose, $\mathbf{x} \in \mathbb{R}^6$ is the 3D motion increment.

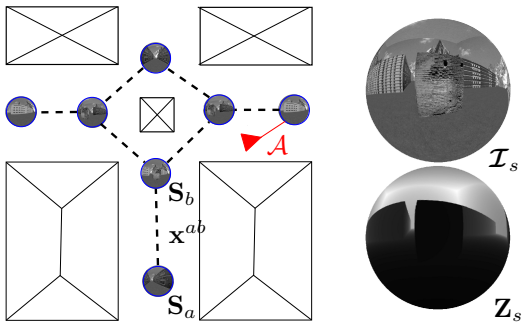


Figure 3. Ego-centric representation: graph of augemented reference images $\mathcal{G}$ containing grey level pixel intensities and their corresponding depths. The reference images are projected onto a unit sphere $\mathcal{S}^2$. $\mathcal{A}$: an agent (robot or person) is shown connected to the graph.

7. Automatic 3D model acquisition

A 3D model is a graph $\mathcal{G} = \{\mathbf{S}_1, \ldots, \mathbf{S}_N\}$ of augmented reference images, each one connected by a 3D pose.

- Reference augmented images are extracted by dense matching from a classic **stereo** rig.
- Pose estimation is performed by **direct** minimization between the last augmented reference image and the current image in the sequence.
- A new reference image is added to the model by monitoring the scale of the minimized error distribution.

The obtained model is directly used for real-time camera localisation.

8. Real-time localisation demonstration

In the demo [2], a dense texture-mapped 3D model will be acquired using an ego-centric based approach. This will involve using multiple key reference images to cover the mapped environment. These reference images are 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.

The 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 [5]. 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).

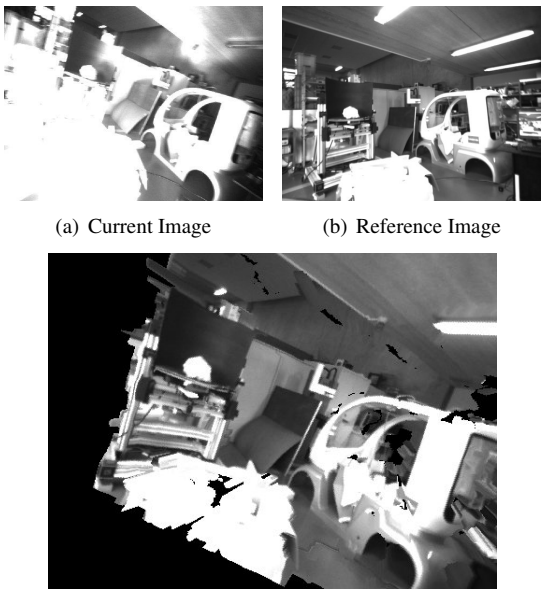
If tracking fails, a recovery procedure finds the closest reference image from the database by performing direct low-resolution tracking between the first image acquired from the live camera and each image from the database. The best score in terms of "Zero Normalised Correlation" is then chosen as the 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.

In summary:

- **45 Hz** monocular (800x600) 3D tracking is achieved on CPU by using only salient pixels in the minimisation process [5].
- The camera pose is estimated using one or multiple augmented reference images extracted from the model.
- The proposed method is able to track all the 6 dof of the camera in various perturbations cases: camera **shaking** and motion **blur**, **defocus** and **aperture changes**, **occlusions**, **saturation** and **shadows**.

The figures shows different results captured from live hand-held tracking. The virtual images (b) are rendered according to the estimated pose using the reference image depth and texture. It can be seen that even if the current image differs a lot form the reference one, the method is still able to track the camera.

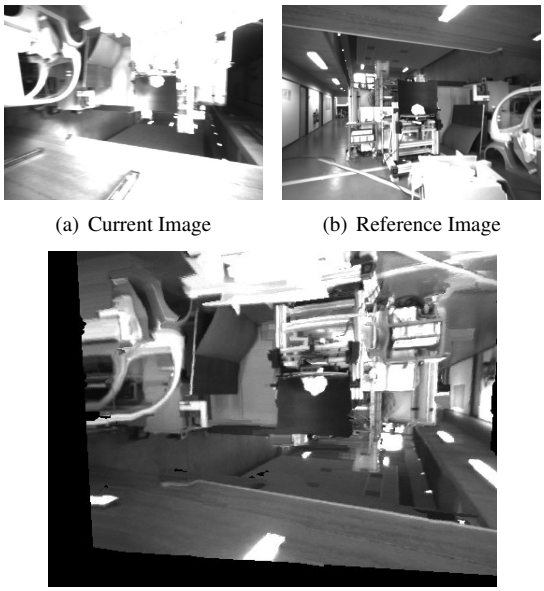
[2] http://www.i3s.unice.fr/~comport/videos/LDRMC11_DenseVisualTrackingLQ.avi



(a) Current Image (b) Reference Image

(c) Virtual Image

Figure 4. Diffuse daylight variation, full 6 dof motion with local saturations



(a) Current Image (b) Reference Image

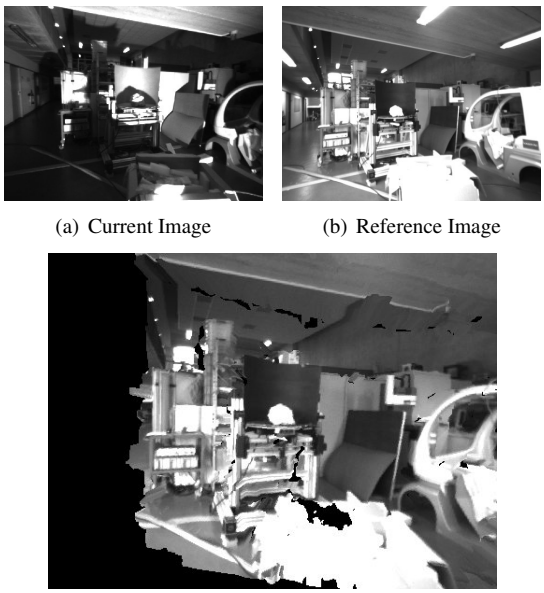
(c) Virtual Image

Figure 5. Diffuse daylight variation, 180 degrees rotation

9. Conclusion

The real-time tracking method presented here will be shown to track robustly in complex 3D scenes at 45Hz on very challenging image sequences. It efficiently combines a model-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 using a stereo camera rig. Salient pixel selection allows 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.



(a) Current Image (b) Reference Image

(c) Virtual Image

Figure 6. Spot illumination with shadows, and local reflections.

References

[1] A. Comport, E. Malis, and P. Rives. Accurate quadrifocal tracking for robust 3d visual odometry. In *IEEE Conference on Robotics and Automation*, pages 40–45, April 2007.

[2] A. 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.

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

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

[5] M. Meilland, A. 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, October 2010.

[6] M. Meilland, A. Comport, and P. Rives. Real-time Dense Visual Tracking under Large Lighting Variations. In *British Machine Vision Conference*, University of Dundee, 29 August - 2 September 2011.