OUTDOOR DELAYED-STATE VISUALLY AUGMENTED ODOMETRY

Viorela Ila, Juan Andrade-Cetto and Alberto Sanfeliu

Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Llorens Artigas 4-6, Barcelona, 08028 Spain.

Abstract: This paper presents an efficient approach to outdoor visually augmented odometry. The technique computes relative pose constraints via a robust least squares minimisation of 3D point correspondences, which are in turn obtained from the matching of SIFT features over two consecutive image pairs. Pose constraints are then used to build a history of pose estimates with and incremental delayed-state information filter. The efficiency of the approach resides on the exact sparseness of the delayed-state information form used.

Keywords: visual odometry, least squares pose estimation, delayed-state SLAM.

1. INTRODUCTION

Accurate localisation is an essential component for any outdoor autonomous navigation system. When no exteroceptive sensors are available, such as GPS signals, a robot must rely on its own sensors to maintain good localisation. Most successful vision-based localisation techniques require the environment to be contaminated with distinctive artificial landmarks. It is desirable however for outdoor navigation, that natural features or salient interest points be used insted. Feature observations generated from different poses can then be matched to produce an estimate of the change in camera pose. When not only the sensor pose is maintained, but the location of the features also, the problem is typically referred as Simultaneous Localisation and Mapping (SLAM) in the robotics community, or Structure From Motion (SFM) in the computer vision community. Recent approaches to SLAM suggest that instead of estimating feature locations, a *delayed-state* history of pose constraints could be maintained (Bosse *et al.*, 2004; Cole and Newman, 2006; Eustice *et al.*, 2006).

In this paper we present such an approach to outdoor vision-based delayed-state SLAM. Given the fact that we are not yet closing large loops and only using consecutive frames, we rather refer to the method only as *visually-augmented odometry.* The technique iterates as follows: SIFT image features are extracted and matched from two stereo image pairs at consecutive frames. In order to obtain a set of 3D feature matches, point correspondences are found in all four images and then, they are independently triangulate in each set of stereo images. These are used to compute a least squares best fit pose transformation. Robust feature outlier rejection is obtained via RANSAC during the computation of the best camera pose constraint. These camera pose constraints are used as relative pose measurements in a delayed-state information-form SLAM. A substantial computational complexity advantage of the delayed-state information-form SLAM is that predictions

This research is supported by the Spanish Ministry of Education and Science under a Juan de la Cierva Postdoctoral Fellow to V. Ila, a Ramón y Cajal Postdoctoral Fellow to J. Andrade-Cetto, the NAVROB Project DPI 2004-5414, and the EU URUS project FP6-IST-045062.

and updates take constant time given its exact sparseness (Eustice *et al.*, 2006).

In a delayed-state SLAM representation, estimation is performed on an information vector of a history of poses, in which the information links relating two nearby poses are updated from sensor reading matches emanating from such two locations. The computed pose difference relating such sensor matches is commonly referred as *pose constraint*, in the sense that it provides information that tightens the link between the two poses, thus reducing their relative localisation uncertainty. Robust pose constraints are typically computed from the matching of 2D range data in planar scenes (Bosse *et al.*, 2004), 3D range data in outdoor scenes (Cole and Newman, 2006), or indoor and outdoor image data (Se *et al.*, 2005; Eustice *et al.*, 2006).

Some approaches that use local maps within the EKF SLAM context to augment dead-reckoning visual data include the Compressed Filter (Guivant and Nebot, 2001) or Postponement (Knight *et al.*, 2001). Others rely only on pure vision instead for the computation of the egomotion for example by tracking Harris features at frame-rate (Nister *et al.*, 2004) (Levin and Szeleski, 2004).

SIFTs have been used for indoor SLAM in the past, most notably in (Se *et al.*, 2002). And also, for closing large loops (Se *et al.*, 2005), as well as for local 3D map alignment. In these works, a conventional EKF SLAM formulation is used. Our interest lies in using such robust features in a delayed-state information-form framework, and for outdoor environments. The difference of our approach when compared to that in (Eustice *et al.*, 2006) is in the use of SIFT feature matches over consecutive pairs of images for the computation of 6D relative pose constraints (3D position and Euler angles) instead of the use of combined Harris and SIFT matches over monocular camera sequences for the computation of 5DOF relative orientation constraints (azimuth, elevation and Euler angles).

2. COMPUTATION OF VISION-BASED POSE CONSTRAINTS

2.1 Feature Extraction

Simple correlation-based features, such as Harris corners (Harris and Stephens, 1988) or Shi and Tomasi features (Shi and Tomasi, 1994), are of common use in vision-based SFM and SLAM. From the early uses of Harris himself to the popular work of Davison (Davison *et al.*, 2007). This kind of features can be robustly tracked when camera displacement is small and are tailored to real-time applications. However, given their sensitivity to scale, their matching is prone to fail under larger camera motions; less to say for loop-closing hypotheses testing. Given their scale and local affine invariance properties, we opt to use SIFTs instead (Lowe, 2004), as they constitute a better option for matching visual features from varying poses. To deal with scale and affine distortions in SIFTs, keypoint patches are selected from difference-of-Gaussian images at various scales, for which the dominant gradient orientation and scale are stored.

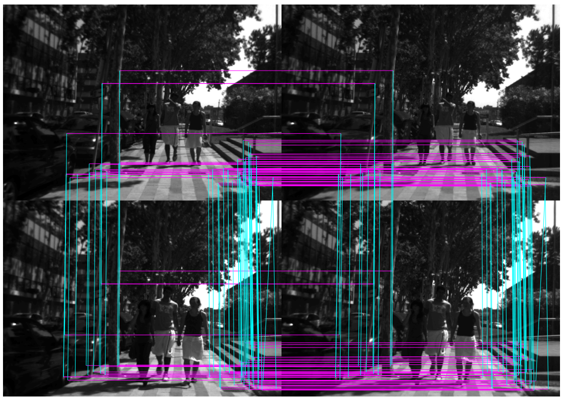


Fig. 1. SIFT correspondences in two consecutive stereo image pairs after outlier removal using RANSAC.

In our system, two consecutive 640×480 image pairs are acquired from a well calibrated stereo rig[1]. Features are extracted and matched in the four image set. The surviving ones are then independently triangulated in each set of stereo images enforcing epipolar and disparity constraints. The epipolar constraint is enforced by allowing feature matches only within ± 1 pixel rows on rectified images. The disparity constraint is set to allow matches within a $1 - 10$ meter range, where camera resolution is best. See Figure 1. The result is a set of two clouds of matching 3D points $\mathbf{p}_t$ and $\mathbf{p}_{t+1}$ referenced to the coordinate frames of the left camera before and after the motion step, respectively.

2.2 Pose Estimation

The homogeneous transformation relating the two aforementioned clouds of points can be computed by solving a set of equations of the form

$$\mathbf{p}_t = \mathbf{R}\mathbf{p}_{t+1} + \mathbf{t}\,. \qquad (1)$$

A solution for the rotation matrix $\mathbf{R}$ is computed by minimising the sum of the squared errors between the rotated directional vectors[2] of feature

[1] Point Gray's Bumblebee firewire stereo camera.

[2] A directional vector $\mathbf{v}$ can be computed as the unit norm direction along $\mathbf{p}$, and indicates the orientation of such point.

matches after the motion step and the corresponding directional vectors prior to the motion step. The solution to this minimisation problem gives an estimate of the orientation of one cloud of points with respect to the other, and can be expressed in quaternion form as

$$\frac{\partial}{\partial \mathbf{R}} \left(\mathbf{q}^\top \mathbf{A} \mathbf{q}\right) = 0\,, \tag{2}$$

where $\mathbf{A}$ is given by

$$\mathbf{A} = \sum_{i=1}^{N} \mathbf{B}_i \mathbf{B}_i^\top\,, \tag{3}$$

$$\mathbf{B}_i = \begin{bmatrix} 0 & -c_x^i & -c_y^i & -c_z^i \\ c_x^i & 0 & b_z^i & -b_y^i \\ c_y^i & -b_z^i & 0 & b_x^i \\ c_z^i & b_y^i & -b_x^i & 0 \end{bmatrix}\,, \tag{4}$$

and

$$\mathbf{b}^i = \mathbf{v}_{t+1}^i + \mathbf{v}_t^i, \quad \mathbf{c}^i = \mathbf{v}_{t+1}^i - \mathbf{v}_t^i \quad . \tag{5}$$

The quaternion $\mathbf{q}$ that minimises the argument of the derivative operator in the differential equation (2) is the smallest eigenvector of the matrix $\mathbf{A}$.[3] Once the rotation matrix $\mathbf{R}$ is computed, we can use again the matched set of points to compute the translation vector $\mathbf{t}$

$$\mathbf{t} = \sum_{i=1}^{N} \mathbf{p}_t^i - \mathbf{R} \sum_{i=1}^{N} \mathbf{p}_{t+1}^i \quad . \tag{6}$$

It might be the case that SIFT matches occur on areas of the scene that experienced motion during the acquisition of the two image stereo pairs. For example, an interest point might appear at an acute angle of a tree leaf shadow, or on a person walking in front of the robot. The corresponding matched 3D points will not represent good fits to the camera motion model, and might introduce large bias to our least squares pose error minimisation. To eliminate such *outliers*, we resort to the use of RANSAC (Fischler and Bolles, 1981). The use of such a robust model fitting technique allows us to preserve the largest number of point matches that at the same time minimise the square sum of the residuals $\|\mathbf{R}\mathbf{p}_{t+1} + \mathbf{t} - \mathbf{p}_t\|$, as shown in Figure 1.

3. VISUALLY AUGMENTED ODOMETRY IN INFORMATION FORM

3.1 Exactly Sparse Delayed-State SLAM

The Extended Kalman Filter SLAM inference has time complexity quadratic in the number of states. As a consequence, the direct application of EKF SLAM is limited to relatively small environments. In contrast, the delayed-state information-form SLAM has been shown to produce exactly sparse information matrices (Eustice *et al.*, 2006), which in our case are tri-block diagonal, linking consecutive measurements each time. This situation allows for constant predictions and updates, with a considerable advantage in terms of computational cost.

The delayed-state information-form SLAM representation consists on estimating a state vector with the history of poses

$$\mathbf{x} = \left[x_t, y_t, \theta_t, \ldots x_1, y_1, \theta_1\right]^\top\,, \tag{7}$$

parameterised as an inverse normal distribution. This representation, dual to the EKF, maintains the information vector and matrix of the state rather than its mean and covariance.

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}^{-1}(\mathbf{x}; \boldsymbol{\eta}, \boldsymbol{\Lambda})\,, \tag{8}$$

where

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} \qquad \text{and} \qquad \boldsymbol{\eta} = \boldsymbol{\Lambda}\boldsymbol{\mu} \quad . \tag{9}$$

Let $\boldsymbol{\mu}_t$ be the state mean for the pose at time t. In a delayed state representation, the map state is simply the history of pose estimates

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_t \\ \vdots \\ \boldsymbol{\mu}_1 \end{bmatrix}\,,$$

and the information vector is

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}_t \\ \vdots \\ \boldsymbol{\eta}_1 \end{bmatrix}\,.$$

From dead-reckoning readings we get the absolute position and orientation x^d, y^d and θ^d at two consecutive time steps t and $t+1$. The relative travelled distance in polar coordinates is given by the length d and angle ψ. Moreover, the relative change of orientation is $\Delta\theta$:

$$d = \sqrt{(x_{t+1}^d - x_t^d)^2 + (y_{t+1}^d - y_t^d)^2} \tag{10}$$

$$\psi = \tan^{-1}\left(\frac{y_{t+1}^d - y_t^d}{x_{t+1}^d - x_t^d}\right) - \theta_t^d \tag{11}$$

$$\Delta\theta = \theta_{t+1}^d - \theta_t^d\,. \tag{12}$$

[3] If we denote this smallest eigenvector by the 4-tuple $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)^\top$, it follows that the rotational angle θ associated with the rotational transform is given by $\theta = 2cos^{-1}(\alpha_1)$, and the axis of rotation would be given by $\hat{\mathbf{a}} = \frac{(\alpha_2, \alpha_3, \alpha_4)^\top}{\sin(\theta/2)}$. Then, it can be shown that the elements of the rotation submatrix $\mathbf{R}$ are related to the orientation parameters $\hat{\mathbf{a}}$ and θ by

$$\mathbf{R} = \begin{bmatrix} a_x^2 + (1 - a_x^2)c_\theta & a_x a_y c'_\theta - a_z s_\theta & a_x a_z c'_\theta + a_y s_\theta \\ a_x a_y c'_\theta + a_z s_\theta & a_y^2 + (1 - a_y^2)c_\theta & a_y a_z c'_\theta - a_x s_\theta \\ a_x a_z c'_\theta - a_y s_\theta & a_y a_z c'_\theta + a_x s_\theta & a_z^2 + (1 - a_z^2)c_\theta \end{bmatrix}\,,$$

where $s_\theta = \sin\theta$, $c_\theta = \cos\theta$, and $c'_\theta = 1 - \cos\theta$ (Kim and Kak, 1991).

$$\begin{bmatrix} \frac{\Delta x\cos\psi+\Delta y\sin\psi}{d} & \frac{-\Delta x\sin\psi+\Delta y\cos\psi}{d} & 0 & \frac{-\Delta x\cos\psi-\Delta y\sin\psi}{d} & \frac{\Delta x\sin\psi-\Delta y\cos\psi}{d} & d\sin\psi \\ \frac{\Delta x\sin\psi-\Delta y\cos\psi}{d} & \frac{\Delta x\cos\psi+\Delta y\sin\psi}{d} & 0 & \frac{-\Delta x\sin\psi+\Delta y\cos\psi}{d} & \frac{-\Delta x\cos\psi-\Delta y\sin\psi}{d} & -d\cos\psi \\ 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix} \qquad (16)$$

The terms in (10-12) form the input $\mathbf{u}_t$ to the model for the prediction of vehicle motion purely from odometry.

$$\begin{aligned} x_{t+1} &= x_t + d\cos(\theta_t + \psi) && (13) \\ y_{t+1} &= y_t + d\sin(\theta_t + \psi) && (14) \\ \theta_{t+1} &= \theta_t + \Delta\theta\,. && (15) \end{aligned}$$

As in most SLAM formulations, white noise $\mathbf{w}_t$ with covariance $\mathbf{Q}$ is added to the vehicle motion prediction model (13-15), and its linearised version used in the computation of covariance prediction (information prediction in our case)

$$\begin{aligned} \mathbf{x}_{t+1} &= f(\mathbf{x}_t, \mathbf{u}_t) + \mathbf{w}_t \\ &\approx f(\boldsymbol{\mu}_t, \mathbf{u}_t) + \mathbf{F}(\mathbf{x}_t - \boldsymbol{\mu}_t) + \mathbf{w}_t \end{aligned} \,. \qquad (17)$$

The revision of the entire history of poses, as a result of adding the odometry information that links the current and predicted poses, can be computed in information form (Eustice *et al.*, 2006) with

$$\bar{\boldsymbol{\eta}} = \begin{bmatrix} \mathbf{Q}^{-1}\left(\mathbf{f}(\boldsymbol{\mu}_t, \mathbf{u}_t) - \mathbf{F}\boldsymbol{\mu}_t\right) \\ \boldsymbol{\eta}_t - \mathbf{F}^\top\mathbf{Q}^{-1}\left(\mathbf{f}(\boldsymbol{\mu}_t, \mathbf{u}_t) - \mathbf{F}\boldsymbol{\mu}_t\right) \\ \boldsymbol{\eta}_{t-1:1} \end{bmatrix}, \qquad (18)$$

and the associated information matrix is

$$\bar{\boldsymbol{\Lambda}} = \begin{bmatrix} \mathbf{Q}^{-1} & -\mathbf{Q}^{-1}\mathbf{F} & \mathbf{0} \\ \mathbf{F}^\top\mathbf{Q}^{-1} & \boldsymbol{\Lambda}_{t,t} + \mathbf{F}^\top\mathbf{Q}^{-1}\mathbf{F} & \boldsymbol{\Lambda}_{t,t-1} \\ \mathbf{0} & \boldsymbol{\Lambda}_{t-1:1,t} & \boldsymbol{\Lambda}_{t-1:1,t-1:1} \end{bmatrix}, \qquad (19)$$

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & -d\sin(\theta_t + \psi) \\ 0 & 1 & d\cos(\theta_t + \psi) \\ 0 & 0 & 1 \end{bmatrix}. \qquad (20)$$

Augmenting the information vector in this form introduces shared information only between the new robot pose $\mathbf{x}_{t+1}$ and the previous one $\mathbf{x}_t$. moreover, the shared information between $\mathbf{x}_{t+1}$ and the delayed-states ($t-1$ to 1) is always zero, resulting in an information matrix with a block tridiagonal structure.

Given the fact that in the information-form measurement updates are additive, they can be also computed in constant-time. Pose differences relating the current and previous poses, as measured by our vision system would be of the form

$$\begin{aligned} \mathbf{z}_{t+1} &= \mathbf{h}(\mathbf{x}_{t+1:t}) + \mathbf{v}_{t+1} \\ &\approx \mathbf{h}(\bar{\boldsymbol{\mu}}_{t+1:t}) + \mathbf{H}(\mathbf{x}_{t+1:t} - \bar{\boldsymbol{\mu}}_{t+1:t}) + \mathbf{v}_{t+1} \end{aligned} \,, \qquad (21)$$

with $\mathbf{v}_{t+1}$ the zero mean, white measurement noise with covariance $\mathbf{R}$ and $\mathbf{H}$ the measurement Jacobian in (16) with

$$\begin{aligned} d &= \sqrt{(x_{t+1} - x_t)^2 + (y_{t+1} - y_t)^2} && (22) \\ \psi &= \tan^{-1}\left(\frac{y_{t+1} - y_t}{x_{t+1} - x_t}\right) - \theta_t && (23) \end{aligned}$$

and

$$\begin{aligned} \Delta x &= x_{t+1} - x_t && (24) \\ \Delta y &= y_{t+1} - y_t\,. && (25) \end{aligned}$$

Thus, our nonlinear measurement model is

$$\begin{aligned} z_{x_{t+1}} &= d\cos(\psi) && (26) \\ z_{y_{t+1}} &= d\sin(\psi) && (27) \\ z_{\theta_{t+1}} &= \theta_{t+1} - \theta_t\,. && (28) \end{aligned}$$

The update to the current and previous entries in the information vector and information matrix become:

$$\boldsymbol{\eta}_{t+1:t} = \bar{\boldsymbol{\eta}}_{t+1:t} + \mathbf{H}^\top\mathbf{R}^{-1}(\mathbf{z}_{t+1} - \mathbf{h}(\bar{\boldsymbol{\mu}}_{t+1:t}) + \mathbf{H}\bar{\boldsymbol{\mu}}_{t+1:t}) \qquad (29)$$

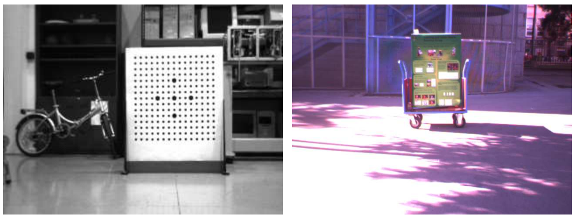
$$\boldsymbol{\Lambda}_{t+1:t,t+1:t} = \bar{\boldsymbol{\Lambda}}_{t+1:t,t+1:t} + \mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}\,. \qquad (30)$$

It is worth stressing that in this visually augmented SLAM respresentation, the Jacobian $\mathbf{H}$ is always sparse (Thrun *et al.*, 2004) and, in consequence only the diagonal blocks of $\boldsymbol{\Lambda}$ will be updated. Non-zero off-diagonal terms would appear only for loop closing situations.

4. EXPERIMENTS

4.1 Vision-based Pose Constraints

An initial experiment was conducted to test the accuracy of our method for the computation of vision-based pose constraints. Our stereo camera was calibrated using the pattern shown in Figure 2(a) using the technique described in (Faugeras, 1993). Given that our calibration technique relies on least squares error minimisation, it was imperative to use a large calibration pattern which was placed at distances of 2, 3, and 4 meters, in order to sample as best as possible the actual workspace of the sensor. Intrinsic and extrinsic parameters of



(a) Calibration pattern (b) Artificial landmark

Fig. 2. Calibration pattern and the artificial landmark placed in an outdoor environment and used to validate the accuracy of stereo reconstruction.

both cameras were obtained, as well as the rigid transformation between them.

Once the camera system was calibrated, a controlled experiment was performed to test the accuracy in stereo reconstruction taking into account the effect of daylight on images and other pervasive artifacts that might contaminate an outdoor scene. To that aim, we located an artificial landmark as shown in Figure 2(b) at a fixed location, and had a synchro drive mobile robot[4] travel through a predefined path taking images at fixed distances of 2, 3, ..., 6 meters from the landmark.

Table 1 shows the average 3D reconstruction error along the x,y and z axes when comparing a set of SIFT features with their manually measured correspondences. The result of the experiment indicates that while our vision-based 3D reconstruction algorithm would have an average error of nearly $2cm$ along the xy plane, depth reconstructions could grow as large as $20cm$. These results suggest the use of a measurement covariance matrix for the SLAM implementation with

$$\mathbf{R} = \begin{bmatrix} (0.02cm)^2 & 0 & 0 \\ 0 & (0.02cm)^2 & 0 \\ 0 & 0 & (\tan^{-1}(2/20))^2 \end{bmatrix} . \tag{31}$$

Table 1. Averrage errors obtained by comparing the manually measured positions and the positions obtained using the vision system.

Dist.(m)	Err X (m)	Err Y(m)	Err Z(m)
2.0	0.01249	0.00831	0.06885
3.0	0.02162	0.01854	0.07128
4.0	0.02314	0.01977	0.09121
5.0	0.02421	0.02243	0.16109
6.0	0.03468	0.02742	0.19377

4.2 Preliminary Experimental Results

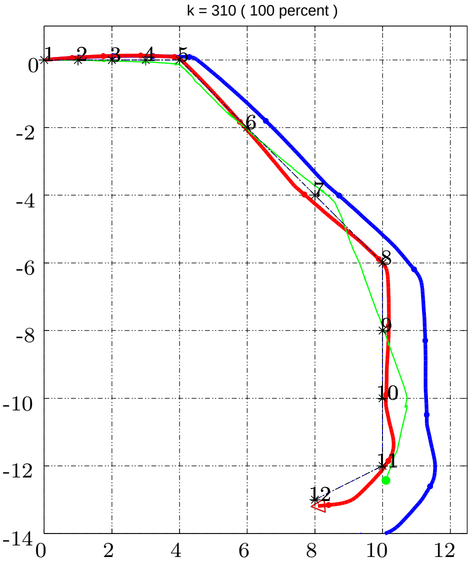


Fig. 3. Odometry-only (blue), vision-only (green), and combined vehicle trajectory (red).

To test our strategy for vision-based augmented odometry we have performed a series of experiments on a urban unstructured environment of small size (approx 200 sq m). A snapshot on one of the tests is shown in Figure 3. The robot was manually driven through a series of predefined via points previously marked on the floor. To record odometry and visual data at such keypoints, the robot stopped at them, only to continue moving after a few seconds. The results of estimating the vehicle motion purely from accumulated raw odometry and purely from concatenating vision pose constraints are shown in the figure as blue and green plots, respectively. The delayed-state information-based revised trajectory resulting from the fusion of the two is shown in red. No motion was accumulated for those cases when not enough SIFT points were obtained during the computation of vision-based pose constraints. This is especially noticeable at the turn at keypoint 8, for which the corresponding turn on accumulated visual odometry happens near keypoint 7. The effect of noninformative vision-based poses at some iterations can be efficiently modelled in our approach only by computing motion predictions from odometry without performing map updates.

Due to the fact that the raw odometry is really poor especially when the vehicle turns and the vision-based pose constraints can fail in translation estimation but provides quite accurate rotation estimation, our SLAM gives more weight to the translation measurements provided by the odometry and to the rotation estimated using SIFT points.

Figure 4 shows the position estimation errors in x and y coordinates relative to the 12 ground truth points marked on the floor. Whereas odometry error accumulates monotonically, vision-based pose constraints vary in accuracy from very accurate up to 2 meters in estimation error. Nontheless the

[4] Activmedia's Pioneer 2DX

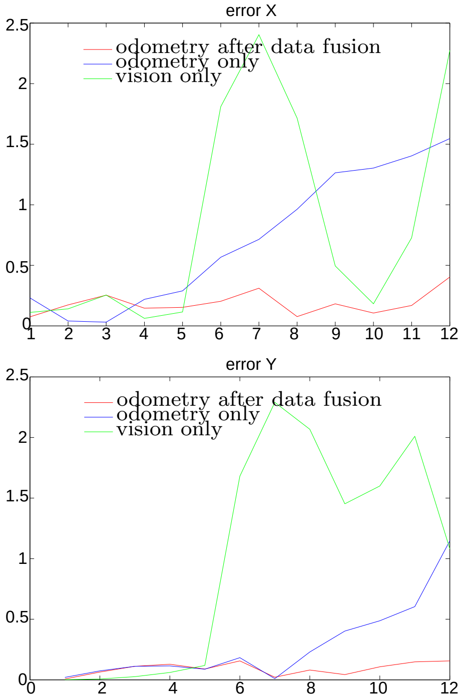


Fig. 4. Pose estimation error considerring 12 ground truth points.

fusion of both provides a consistently revised pose estimate[5] .

5. CONCLUSIONS

This paper proposed an efficient approach to outdoor visual augmented odometry based on an exactly sparse delayed-state filter that uses SIFT features to compute pose constraints. Another type of features that we seek to explore in the future are *Speed Up Robust Features* SURFs (Bay *et al.*, 2006). These features have similar response properties to SIFTs, replacing Gaussian convolutions with Haar convolutions, and a significant reduction in computational cost.

We can conclude saying that, concerning the accuracy of the pose estimation, our approach performs well when comparing the estimated trajectory with ground truth points, and considerably reduces the memory and execution time by using a tri-block diagonal information matrix to link consecutive measurements each time (time and memory increase linearly compared to the quadratic cost of the traditional EKF).

The experimental results presented here constitute a preliminary study on the use of 6D SIFT-based pose transforms for outdoor mapping. We are currently working in vision-based hypothesis testing for loop closure, and expect to report in the near future results of mapping of much larger areas, in the order of 300×300 sq. meters.

REFERENCES

Bay, H., T. Tuytelaars and L. Van Gool (2006). SURF: Speeded up robust features. In: *Proc. 9th European Conf. Comput. Vision*. Vol. 3951 of *Lect. Notes Comput. Sci.*. Springer-Verlag. Graz. pp. 404–417.

Bosse, M., P. Newman, J. Leonard and S. Teller (2004). Simultaneous localization and map building in large-scale cyclic environments using the *atlas* framework. *Int. J. Robot. Res.* **23**(12), 1113–1139.

Cole, D.M. and P.M. Newman (2006). 3D SLAM in outdoor environments. In: *Proc. IEEE Int. Conf. Robot. Automat.*. Orlando. pp. 1556–1563.

Davison, A. J., I.D. Reid, N.D. Molton and O. Stasse (2007). MonoSLAM: Real-time single camera SLAM. *IEEE Trans. Pattern Anal. Machine Intell.* To appear.

Eustice, R.M., H. Singh and J.J. Leonard (2006). Exactly sparse delayed-state filters for view-based SLAM. *IEEE Trans. Robot.* **22**(6), 1100–1114.

Faugeras, O. (1993). *Three-Dimensional Computer Vision. A Geometric Viewpoint*. The MIT Press. Cambridge.

Fischler, M. and R. Bolles (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM* **24**, 381–385.

Guivant, J. E. and E. M. Nebot (2001). Optimization of simultaneous localization and map-builidng algorithm for real-time implementation. *IEEE Trans. Robot. Automat.* **17**(3), 242–257.

Harris, C. G. and M. Stephens (1988). A combined corner edge detector. In: *Proc. Alvey Vision Conf.*. Manchester. pp. 189–192.

Kim, W. Y. and A. C. Kak (1991). 3D object recognition using bipartite matching embedded in discrete relaxation. *IEEE Trans. Pattern Anal. Machine Intell.* **13**(3), 224–251.

Knight, J., A. Davision and I. Reid (2001). Towards constant time SLAM using postponement. In: *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*. Vol. 1. Maui. pp. 405–413.

Levin, A. and R. Szeleski (2004). Visual odometry and map correlation. In: *Proc. 18th IEEE Conf. Comput. Vision Pattern Recog.*. Washington. pp. 611–618.

Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision* **60**(2), 91–110.

Nister, D., O. Naroditsky and J. Bergen (2004). Visual odometry. In: *Proc. 18th IEEE Conf. Comput. Vision Pattern Recog.*. Washington. pp. 652–659.

Se, S., D. Lowe and J. Little (2002). Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *Int. J. Robot. Res.* **21**(8), 735–758.

Se, S., D. Lowe and J. Little (2005). Vision-based global localization and mapping for mobile robots. *IEEE Trans. Robot.* **21**(3), 364–375.

Shi, J. and C. Tomasi (1994). Good features to track. In: *Proc. 9th IEEE Conf. Comput. Vision Pattern Recog.*. Seattle. pp. 593–600.

Thrun, S., Y. Liu, D. Koller, A. Y. Ng, Z. Ghahramani and H. Durrant-Whyte (2004). Simultaneous localization and mapping with sparse extended information filters. *Int. J. Robot. Res.* **23**(7-8), 693–716.

[5] We must take into consideration that placing the centre of the robot to an exact ground point during the experiment is not an easy task and might introduce some error in these plots. The comparison only shows the empirical result that a sparse delayed-state information filter for the fusion of odometry and vision might be a good strategy for large scale outdoor SLAM