IAV2004 - PREPRINTS
5th IFAC/EURON Symposium on Intelligent Autonomous Vehicles
Instituto Superior Técnico, Lisboa, Portugal
July 5-7, 2004

REAL-TIME 3D SLAM WITH WIDE-ANGLE VISION

Andrew J. Davison [*,1] **Yolanda González Cid** [**]
Nobuyuki Kita [***]

[*] *Robotics Research Group, Dept. of Engineering Science,*
University of Oxford, Oxford OX1 3PJ, UK
[**] *Dpto. de Ciencias, Matemáticas e Informática,*
Universidad de las Islas Baleares, 07071 Palma de Mallorca, Spain
[***] *Intelligent Systems Institute, AIST Tsukuba Central 2,*
1-1 Umezono, Tsukuba, Japan

Abstract: The performance of single-camera SLAM is improved when wide-angle optics provide a field of view greater than the 40 to 50 degrees lenses normally used in computer vision. The issue is one of feature contact: each landmark object mapped remains visible through a larger range of camera motion, meaning that feature density can be reduced and camera movement range can be increased. Further, localisation stability is improved since features at widely differing viewing angles are simultaneously visible. We present the first real-time (30 frames per second), fully automatic implementation of 3D SLAM using a hand-waved wide-angle camera, and demonstrate significant advances in the range and agility of motions which can be tracked over previous narrow field-of-view implementations.

Keywords: SLAM, wide-angle vision, real-time

1. INTRODUCTION

Real-time Simultaneous Localisation and Mapping (SLAM) will be a key component of any autonomous robot system. Successful implementations of SLAM have generally been achieved with laser, sonar or stereo vision range sensors and built maps for controlled robots moving in 2D (e.g. Davison and Murray (1998); Newman et al. (2002); Thrun et al. (2000); Castellanos (1998)). Recent research however has proved that real-time 3D SLAM can be achieved using monocular vision as the only sensory input and using only weak motion modelling (Davison (2003); Mayol et al. (2003)) — indicating not only that vision will become increasingly important as a cheap, compact and flexible tool for robot navigation but that visual SLAM will be able to play a role in other domains in which automatic localisation is required, such as wearable computing, human-computer interface and television.

This research in visual SLAM has much in common with the wealth of research in the Structure from Motion (SFM) field in computer vision over the past two decades, which has produced impressive automatic systems for camera trajectory recovery and 3D structure computation (e.g. Fitzgibbon and Zisserman (1998); 2d3). The

[1] This work has been partially supported by CICYT DPI2001-2311-C03-02, FEDER funding and a grant co-financed by the Universitat de les Illes Balears and the European Social Fund. AJD is supported by an EPSRC Advanced Research Fellowship. Thank you to Ben Tordoff and the Active Vision Laboratory, Oxford, for camera calibration software.

approach taken in SFM however has generally been very different from SLAM because the applications aimed at have not required real-time operation, and trajectory and structure computation are able to proceed off-line. Although some real-time SFM systems (e.g. Nistér (2003)) have been produced by efficient implementation of frame-to-frame SFM steps, for real-time operation in which repeatable localisation is possible (and motion uncertainty does not grow without bound over time) the sequential SLAM approach familiar from mobile robotics is essential due to the fundamental emphasis placed on propagation of uncertainty.

In this paper we extend the single camera SLAM work of Davison by replacing the standard perspective camera used with one which has a wider field of view of over 90°, and demonstrate significantly improved SLAM results, with increased movement range, accuracy and ability to track agile motion. Wide-angle lenses typically do not fit standard perspective projection models and we give details of the perspective + distortion model implemented for accurate camera calibration.

That wide field-of-view sensors are beneficial in localisation and mapping is something that has been clear to many researchers. The laser range-finders from Sick commonly used in robot SLAM provide a field of view of 180°, and similar ranges have been provided by active steerable vision (Davison and Murray (1998)) or scanning sonar(Manyika and Durrant-Whyte (1993)). In computer vision, recently there has been emphasis on catadioptric cameras which use a camera/mirror system to provide omni-directional imaging, and work on off-line structure and motion (Geyer and Daniilidis (2001)) and instantaneous motion estimation using these. In particularly impressive work, Bosse et al. (2002) used an omni-directional camera to detect the vanishing points of lines as part of a laser-based SLAM system.

While omni-directional sensors are certainly appealing from a theoretical point of view, we prefer in this work to look at simpler, compact optics to provide the widest range of application opportunities. The lens used in this paper is a low-cost wide-angle unit with volume less than 1cm^3.

2. SINGLE CAMERA SLAM

Davison (2003) demonstrated SLAM at 30Hz for a camera waved in the hand, building on-line a sparse map of features to serve as localisation landmarks. Here we summarise this work before going on to explain the changes necessary to incorporate a wide-angle lens.

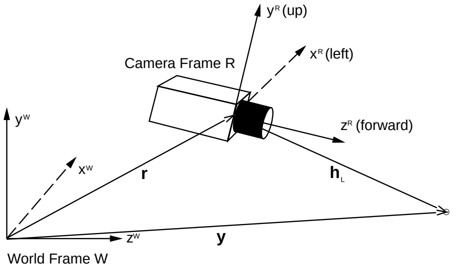


Fig. 1. Frames and vectors in camera geometry.

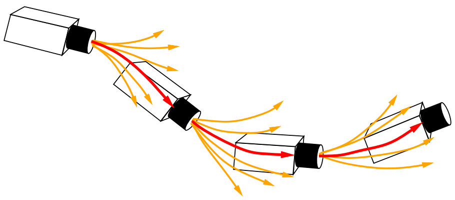


Fig. 2. Visualisation of the "constant velocity" model for smooth motion.

A full-covariance Extended Kalman Filter (EKF) approach is used, storing the estimated state and covariance of the system at any instant as follows:

$$\hat{\mathbf{x}} = \begin{pmatrix} \hat{\mathbf{x}}_v \\ \hat{\mathbf{y}}_1 \\ \hat{\mathbf{y}}_2 \\ \vdots \end{pmatrix} , \ \mathsf{P} = \begin{bmatrix} \mathsf{P}_{xx} & \mathsf{P}_{xy_1} & \mathsf{P}_{xy_2} & \dots \\ \mathsf{P}_{y_1x} & \mathsf{P}_{y_1y_1} & \mathsf{P}_{y_1y_2} & \dots \\ \mathsf{P}_{y_2x} & \mathsf{P}_{y_2y_1} & \mathsf{P}_{y_2y_2} & \dots \\ \vdots & \vdots & \vdots & \end{bmatrix} . \quad (1)$$

Explicitly, the camera's state vector $\mathbf{x}_v$ comprises a metric 3D position vector $\mathbf{r}^W$, orientation quaternion $\mathbf{q}^{RW}$, velocity vector $\mathbf{v}^W$ and angular velocity vector ω^W relative to a fixed world frame (13 parameters — the use of a quaternion for orientation is non-minimal, but preferred for reasons of conditioning and ease of manipulation). Feature states $\mathbf{y}_i$ are 3D position vectors. Figure 1 defines coordinate frames and vectors.

With the aim of flexible application, it is assumed that odometry is not available, and in the EKF prediction step a model for smooth motion anticipates Gaussian-distributed perturbations $\mathbf{V}^W$ and $\mathbf{\Omega}^W$ to the camera's linear and angular velocity at each time-step — modelling motion with a generally smooth character. The explicit process model for motion in a time-step Δt is:

$$\mathbf{f}_v = \begin{pmatrix} \mathbf{r}^W_{new} \\ \mathbf{q}^{WR}_{new} \\ \mathbf{v}^W_{new} \\ \omega^W_{new} \end{pmatrix} = \begin{pmatrix} \mathbf{r}^W + (\mathbf{v}^W + \mathbf{V}^W)\Delta t \\ \mathbf{q}^{WR} \times \mathbf{q}((\omega^W + \mathbf{\Omega}^W)\Delta t) \\ \mathbf{v}^W + \mathbf{V}^W \\ \omega^W + \mathbf{\Omega}^W \end{pmatrix} (2)$$

Figure 2 illustrates how this models potential deviations from a constant velocity trajectory. Implementation requires calculation of the Jacobians of this process function with respect to both $\mathbf{x}_v$ and the perturbation vector (not presented here).

The features used in the map are natural points of high image interest detected using the operator of Shi and Tomasi (1994) and saved as square image template patches. Figure 3 shows the type of image regions typically detected, corresponding mainly to corners or well-localised small objects. When a feature is first initialised, measurement from a single camera position provides good information on its direction relative to the camera, but its depth is initially unknown beyond potentially very weak prior information on the typical depths of objects in the scene. A semi-infinite 3D line is therefore initialised into the SLAM map, with end-point at the camera optical centre and direction derived from the image measurement: the 3D location of the feature lies somewhere along this line. The parameters describing the line have Gaussian-distributed uncertainties and corresponding entries in the SLAM covariance matrix, but to represent the non-Gaussian uncertainty in depth a discrete particle probability distribution is initialised along this coordinate with an initial flat profile representing complete uncertainty. As the camera moves and subsequent images are acquired, each particle hypothesis for depth is repeatedly tested and their probabilities evolve. Figure 3(b) illustrates image search in a set of overlapping ellipses corresponding to the particles, and (c) the progression of the depth PDF from flat to a final peak at which point it can be replaced with a Gaussian and the feature fully initialised as a 3D point in the SLAM map. This process can take from 2–10 frames depending on the camera motion and uncertainty.

Figure 3(d) illustrates active search for fully-initialised features during normal operation. The uncertainty in the relative position of the camera and features is projected into the current image and used to deduce elliptical search regions corresponding to 3 standard deviation confidence intervals within which the features are known to lie with high probability. Expensive normalised correlation search for matches can be restricted to these regions and this gives the algorithm the efficiency necessary for real-time implementation.

Note that the 3D positions and image descriptions of a small number of features (the four corners of an A4 piece of paper are enough) are required to bootstrap the SLAM system, principally to provide information on the overall metric scale of the map and camera motion. All other features are detected automatically and the initialisation target can soon move out of the field of view or even be removed. Heuristic map-management criteria are used to decide when to initialise new features: essentially, the requirement is to keep a pre-defined number of features visible from all camera locations. A typical number used is 10; whenever fewer than 10 features are visible new ones are detected and initialised. Importantly, features are not deleted from the map when the leave the field of view, but remain in the map and can be re-observed when the camera moves back and they become visible again. In some cases it is necessary to delete features which are not being reliably matched on a regular basis: some features detected will be frequently occluded or may contain parts of objects at very different depths. These features will lead to failed correlation attempts and can be removed from the map automatically.

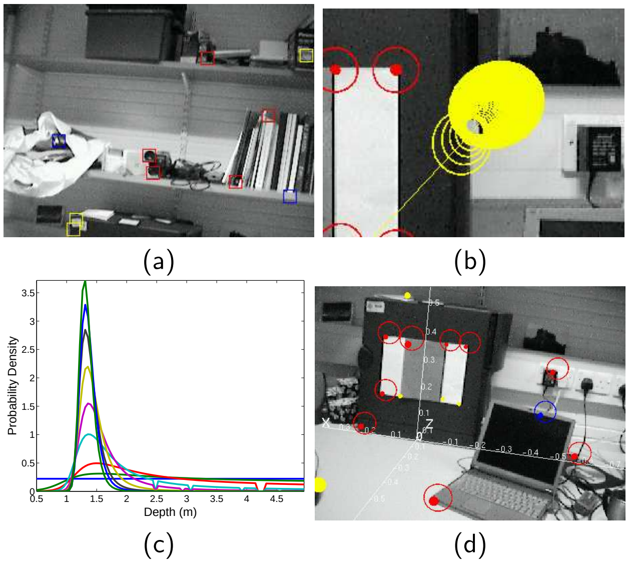


Fig. 3. Feature detection, initialisation and matching: (a) 11 × 11 pixel patches detected a features; (b) searching a set of hypotheses for feature depth which project as image search ellipses; (c) probability distribution over depth refined to a peak over several time-steps; (d) elliptical search regions for mapped features during normal operation.

The main limitation in the results of Davison (2003) is that the range of motion of the camera for which tracking can be maintained is limited to the desk-top scale, due largely to computational restrictions on the number of mapped feature estimates which can be maintained with 30Hz updates. Noticeable “jitter” in the camera localisation results is observed, and the speed of camera motion is also limited by the need to initialise new features while maintaining a good overlap with already mapped features within the field of view. A significant factor in these issues is the narrow field of view of the camera used: the features which can be seen simultaneously are fundamentally close together, and this leads to high uncertainty in the camera position estimate attainable. High ambiguity between rotation and translation is typical, especially when the features observed have a small depth range, a common situation. Further, features must be mapped very densely in order that a sufficient number for localisation can be seen at all times. In the following section we describe the steps necessary to incorporate a wide-angle lens into the system.

3. PROJECTION MODEL FOR WIDE-ANGLE LENS

The lens used in this work was the integral lens of the wide-angle version of the Unibrain Fire-i IEEE1394 web-cam module which provides a field of view of around 90° horizontal, 70° vertical. As Figure 4 shows, images from this camera exhibit a large discrepancy from the characteristics of pure pinhole perspective projection, in which straight lines in the 3D scene always project to straight image lines. With narrow field of view lenses, perspective projection is generally a good approximation to the true imaging characteristics, but clearly that is no longer the case here. Instead we assume that the projection can be modelled as the combination of a perspective projection and a radial image distortion in sequence.

From Figure 1, given estimates of the 3D positions of the camera and a feature, the position of the feature relative to the camera is expected to be:

$$\mathbf{h}_L^R = \mathtt{R}^{RW}(\mathbf{y}_i^W - \mathbf{r}^W) \ .$$

Considering first the perspective stage of projection, the position (u, v) at which the feature is expected to be found in the image is found using the pinhole camera model:

$$\mathbf{h}_i = \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u_0 - fk_u \frac{h_{Lx}^R}{h_{Lz}^R} \\ v_0 - fk_v \frac{h_{Ly}^R}{h_{Lz}^R} \end{pmatrix} ,$$

where (using standard computer vision notation) fk_u and fk_v represent the focal length in horizontal and vertical pixel units respectively, and u_0 and v_0 are the coordinates of the principal point.

A radial distortion then warps the perspective-projected coordinates $\mathbf{u}$ to final image position $\mathbf{u}_d$:

$$\mathbf{u} = \begin{pmatrix} u \\ v \end{pmatrix} \quad , \quad \mathbf{u}_d = \begin{pmatrix} u_d \\ v_d \end{pmatrix} \ . \tag{3}$$

The following radial distortion model was chosen because to a good approximation it is invertible Swaminathan and Nayar (2000):

$$u_d - u_0 = \frac{u - u_0}{\sqrt{1 + 2K_1 r^2}} \tag{4}$$

$$v_d - v_0 = \frac{v - v_0}{\sqrt{1 + 2K_1 r^2}} \ , \tag{5}$$

where

$$r = \sqrt{(u - u_0)^2 + (v - v_0)^2} \ . \tag{6}$$

The inverse of this distortion is as follows:

$$u - u_0 = \frac{u_d - u_0}{\sqrt{1 - 2K_1 r_d^2}} \tag{7}$$

$$v - v_0 = \frac{v_d - v_0}{\sqrt{1 - 2K_1 r_d^2}} \ , \tag{8}$$

where

$$r_d = \sqrt{(u_d - u_0)^2 + (v_d - v_0)^2} \ . \tag{9}$$

For implementation as the measurement step of the EKF, the Jacobians of the projection function with respect to camera and feature positions must be found. The Jacobians of the perspective part of the projection are trivial and given elsewhere; forward and backward Jacobians of the distortion function are shown in Table 1.

The camera was calibrated using standard software and a calibration grid, obtaining values $fk_u = fk_v = 195$ pixels, $(u_0, v_0) = (162, 125)$, $K_1 = 6 \times 10^{-6}$ for capture at 320×240 resolution.

4. RESULTS

To demonstrate clearly the advantages provided by the wide-angle lens, an experiment was devised in which two cameras, one with a narrow FOV and one with the wide-angle lens, were mounted rigidly together, side-by-side and parallel, and each connected to a different PC for processing as the rig was waved by hand from a starting position in front of an initialisation target with four known features. All parameters of the two systems were identical (including the motion model noise parameters) apart from camera projection models.

Snapshot results are displayed in Figure 4. The advantages clear in the wide-angle case were:

(1) Better camera motion estimation, and in particular improved disambiguation of rotation and translational movements, as highlighted in Figure 4. In the wide angle camera, features in highly different directions are simultaneously visible, whereas in the narrow view often all the features measured lie very close together both in direction and depth. In such situations small rotations and translations are ambiguous.

(2) Increased movement range: due both to the more efficient sparse mapping possible, and lower rate of increase in uncertainty with motion since old features are visible for longer.

(3) Increased movement accelerations trackable: even adjusting the motion model noise parameters to rather uncertain values (up to 10ms^{-2} in the linear acceleration component) does not mean that the image search regions in the wide angle camera's "zoomed out"

$$\frac{\partial \mathbf{u}_d}{\partial \mathbf{u}} = \begin{bmatrix} \frac{\partial u_d}{\partial u} & \frac{\partial u_d}{\partial v} \\ \frac{\partial v_d}{\partial u} & \frac{\partial v_d}{\partial v} \end{bmatrix} = \begin{bmatrix} \frac{-2(u-u_0)^2 K_1}{(1+2K_1r^2)^{\frac{3}{2}}} + \frac{1}{\sqrt{1+2K_1r^2}} & \frac{-2(u-u_0)(v-v_0)K_1}{(1+2K_1r^2)^{\frac{3}{2}}} \\ \frac{-2(v-v_0)(u-u_0)K_1}{(1+2K_1r^2)^{\frac{3}{2}}} & \frac{-2(v-v_0)^2 K_1}{(1+2K_1r^2)^{\frac{3}{2}}} + \frac{1}{\sqrt{1+2K_1r^2}} \end{bmatrix}$$

$$\frac{\partial \mathbf{u}}{\partial \mathbf{u}_d} = \begin{bmatrix} \frac{\partial u}{\partial u_d} & \frac{\partial u}{\partial v_d} \\ \frac{\partial v}{\partial u_d} & \frac{\partial v}{\partial v_d} \end{bmatrix} = \begin{bmatrix} \frac{2(u_d-u_0)^2 K_1}{(1-2K_1r_d^2)^{\frac{3}{2}}} + \frac{1}{\sqrt{1-2K_1r_d^2}} & \frac{2(u_d-u_0)(v_d-v_0)K_1}{(1-2K_1r_d^2)^{\frac{3}{2}}} \\ \frac{2(v_d-v_0)(u_d-u_0)K_1}{(1-2K_1r_d^2)^{\frac{3}{2}}} & \frac{2(v_d-v_0)^2 K_1}{(1-2K_1r_d^2)^{\frac{3}{2}}} + \frac{1}{\sqrt{1-2K_1r_d^2}} \end{bmatrix}$$

Table 1. Jacobian matrices for the forward and backward distortion models.

view do not grow too large. Large motions appear in general much less abrupt when viewed from the wide-angle camera.

A limitation that remains currently is that tracking can only be maintained while the camera moves in such a way as to stay roughly horizontal, since feature matching takes place through 2D template matching and features that rotate in the image cannot be matched. In current work we are aiming to rectify this by considering features as 3D planar entities in the world, aiming to track motions where cameras can turn upside-down.

5. CONCLUSIONS

Vision is the sensing modality most likely to enable Simultaneous Localisation and Mapping (SLAM) algorithms to be implemented widely in the domestic robots of the future and other various compact devices — camera modules of the type installed in mobile telephones and other devices are now omnipresent, cheap and compact. In this paper we have shown that thoughtful choice of a camera's optical characteristics provides another important advance in the robustness and range of application of single camera SLAM with no increase in computational cost or system complexity. In the future it will be interesting to experiment futher with optical configurations and determine which provides the best performance.

REFERENCES

2d3. 2d3 web based literature. URL http://www.2d3.com/, 2004.

M. Bosse, R. Rikoski, J. Leonard, and S. Teller. Vanishing points and 3d lines from omnidirectional video. In *IEEE International Conference on Image Processing*, 2002.

J. A. Castellanos. *Mobile Robot Localization and Map Building: A Multisensor Fusion Approach.* PhD thesis, Universidad de Zaragoza, Spain, 1998.

A. J. Davison. Real-time simultaneous localisation and mapping with a single camera. In *Proceedings of the 9th International Conference on Computer Vision, Nice*, 2003.

A. J. Davison and D. W. Murray. Mobile robot localisation using active vision. In *Proceedings of the 5th European Conference on Computer Vision, Freiburg*, pages 809–825, 1998.

A. W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. In *Proc. European Conference on Computer Vision*, pages 311–326. Springer-Verlag, June 1998.

C. Geyer and K. Daniilidis. Structure and motion from uncalibrated catadioptric views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Kauai*, 2001.

J. Manyika and H. F. Durrant-Whyte. *Data Fusion and Sensor Management: A Decentralized Information-Theoretic Approach.* Prentice Hall, 1993.

W. W. Mayol, A. J. Davison, B. Tordoff, and D. W. Murray. Applying active vision and SLAM to wearables. In *International Symposium on Robotics Research, Siena, Italy*, 2003.

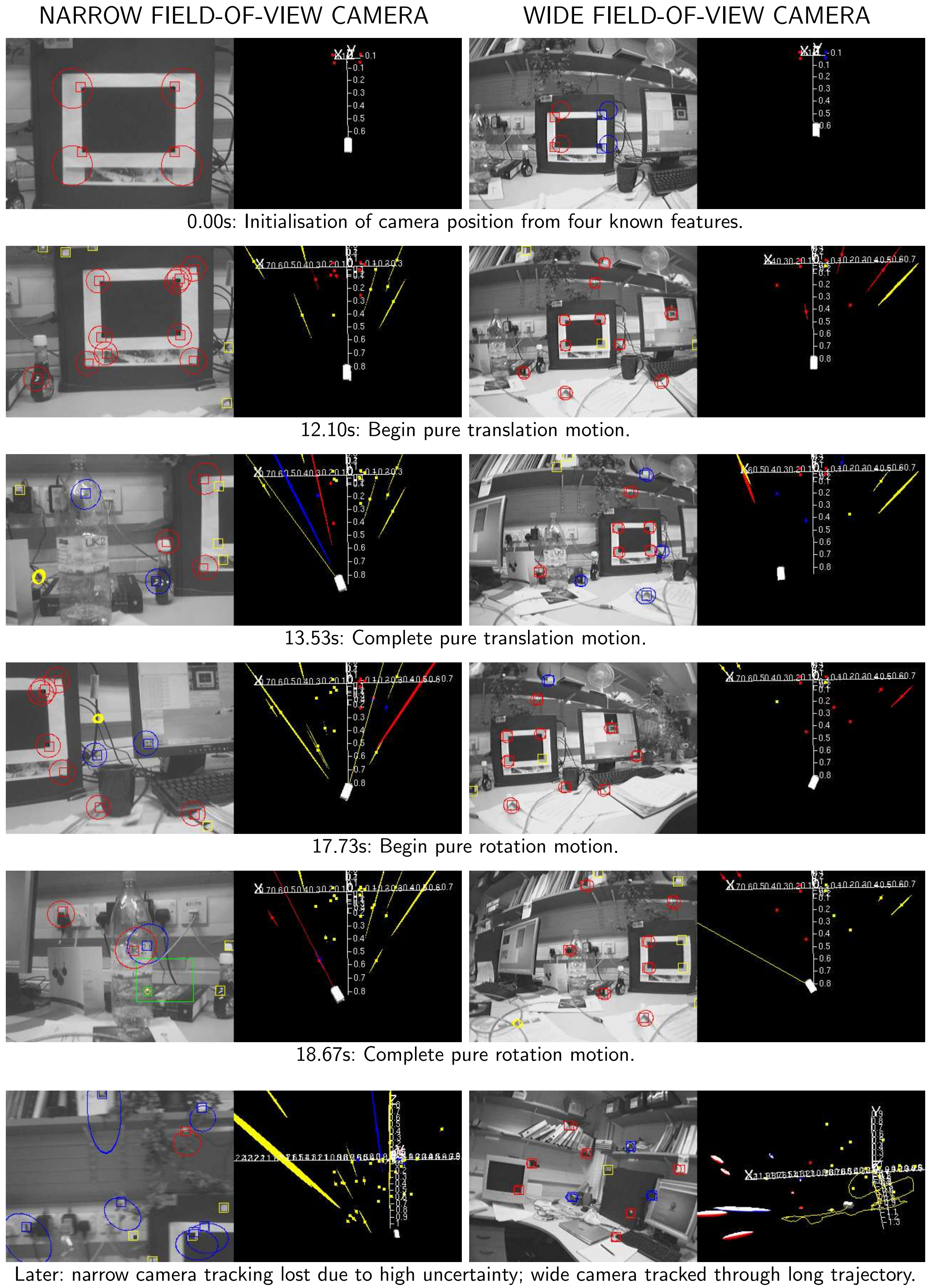
P. M. Newman, J. J. Leonard, J. Neira, and J. Tardós. Explore and return: Experimental validation of real time concurrent mapping and localization. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1802–1809, 2002.

D. Nistér. Preemptive RANSAC for live structure and motion estimation. In *Proceedings of the 9th International Conference on Computer Vision, Nice*, 2003.

J. Shi and C. Tomasi. Good features to track. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

Rahul Swaminathan and Shree K. Nayar. Nonmetric calibration of wide-angle lenses and polycameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10): 1172–1178, 2000.

S. Thrun, W. Burgard, and D. Fox. A real-time algorithm for mobile robot mapping with applications to multi-robot and 3D mapping. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2000.


NARROW FIELD-OF-VIEW CAMERA

WIDE FIELD-OF-VIEW CAMERA

0.00s: Initialisation of camera position from four known features.

12.10s: Begin pure translation motion.

13.53s: Complete pure translation motion.

17.73s: Begin pure rotation motion.

18.67s: Complete pure rotation motion.

Later: narrow camera tracking lost due to high uncertainty; wide camera tracked through long trajectory.

Fig. 4. Synchronised snap-shot results from SLAM processing performed on image sequences obtained from narrow and wide field-of-view cameras rigidly joined side-by-side and waved in the hand, demonstrating improved disambiguation of rotation and translation in the wide case as specific movements are performed. Each snap-shot shows the current image view with overlaid square feature patches and elliptical feature search ellipses together with an external 3D view of the current estimated positions of camera and features (with feature uncertainty ellipsoids). Feature colour-coding is as follows: red = successfully observed at this time-step; blue = attemted measurement failed; yellow = not selected for measurement. The final row shows a typical later situation, where the narrow FOV camera has become lost while the wide camera is tracked through extended motions (the yellow line is its estimated trajectory). Video illustrating the wide-angle SLAM results is available from `http://www.robots.ox.ac.uk/~ajd/Movies/uniwide.mpg`.