Recovering stable scale in monocular SLAM using object-supplemented bundle adjustment

Duncan Frost, Victor A Prisacariu, David W Murray, *Member, IEEE*

***Abstract*—Without knowledge of the absolute baseline between images the scale of a map from a single-camera simultaneous localisation and mapping system is subject to calamitous drift over time. We describe a monocular approach that in addition to point measurements also considers object detections to resolve this scale ambiguity and drift. By placing an expectation on the size of the objects, the scale estimation can be seamlessly integrated into a bundle adjustment. When object observations are available, the local scale of the map is then determined jointly with the camera pose in local adjustments. Unlike many previous visual odometry methods, our approach does not impose restrictions such as constant camera height or planar roadways, and is therefore more widely applicable. We evaluate our approach on the KITTI dataset and show that it reduces scale drift over long-range outdoor sequences with a total length of 40 km. As the scale of objects is known absolutely, metric accuracy is obtained for all sequences. Qualitative evaluation is also performed on video footage from a hand-held camera.**

***Index Terms*—simultaneous localisation and mapping, monocular vision, scale drift, object recognition, bundle adjustment**

I. Introduction

The joint dependencies of sensor measurements on both sensor pose and scene structure — all unknown *a priori* — ensure that simultaneous localisation and mapping algorithms accumulate error as the sensor explores away from its starting position. When using direction and range measurements from, say, LiDAR, RGBD, or stereo cameras (*e.g.* [1]–[3]) this has little impact on the ability to continue exploration. But when using direction-only measurements from a single camera the error is compounded with drift from the depth/speed scaling ambiguity and can incapacitate data collection entirely. For example, keyframe-based approaches often select keyframes based on their proximity to their nearest neighbour. If the estimate of speed is too large, keyframes will be created too rapidly, and if too small they are not created at all. Again, while all methods are able to reduce and redistribute accumulated error upon loop closure (*e.g.* [4]–[8]), drift in monocular methods can be so severe that searching for previously observed landmarks in the map is simply not feasible, notwithstanding progress in the use of geometric and appearance cues [9].

Human experience indicates that scale drift does not occur routinely in our visual systems, even if we act as cyclopean observers [10]. Unless measures are taken to fool us, we have access to other visual, auditory and proprioceptive inputs which, combined with our wealth of prior knowledge, are sufficient to stabilise scale. A key expectation is that the typical size of objects in a class neither shrinks nor expands over short temporal and spatial scales. This paper reports work which imbues monocular SLAM with that same prior. It is shown that with a suitable representation of point landmarks and objects, a modified bundle adjustment can counter scale drift effectively by detecting and tracking instances of object classes of known size distribution and including them within the map. As a continuous estimate of scale is crucial to long term performance, we favour correcting drift frequently in local adjustments rather than infrequently in more global optimisations.

The additional computation is small compared with landmark-only bundle adjustment and it relies on very modest additional information. It could therefore be readily incorporated into any complete pipeline based on adjustment: [11] is an obvious current example. Even in cases where a robotic application has no need to limit its sensing package, either in terms of size or power consumption, it may be of value to squeeze additional information from monocular vision at such modest cost, even if only to provide a validation gate or a prior on other sensor data.

The principal approach offered here is developed in Section IV, where we detail a method of object-supplemented bundle adjustment, one which offers a combined representation for point landmarks and extended objects. Section V outlines its implementation and Section VI gives results both from experimentation using the KITTI dataset and from sequences from handheld phone cameras. Concluding remarks are made in Section VII.

Before presenting the main contribution, Section II considers related methods for correcting scale drift and incorporating objects into SLAM, and Section III briefly reviews our early and unsatisfactory method of using information on object size. Its failure offers insight into a representational cul-de-sac.

II. Related work

Recent accounts of the states of the art in SLAM in general and monocular SLAM in particular are found in [12] and [11] respectively, and here we are concerned only with previous work in mitigating scale uncertainty. Approaches fall into one of two categories: either information from additional sensors is utilised or, as in the method proposed here, assumptions are made about the camera's environment.

In the first category, adding a second camera is the obvious extension (*e.g.* [3], [13]), immediately disambiguating scale if the camera baseline is known. Avoiding a second camera, a

The authors are with the Active Vision Laboratory, Department of Engineering Science, University of Oxford, OX1 3PJ, UK.
E-mail: duncan,victor,dwm@robots.ox.ac.uk.

favoured non-visual sensor is the inertial measurement unit, able to measure acceleration and orientation. As examples, Jung and Taylor [14] used IMU measurements in a structure from motion method. Frames were clustered into temporal windows, with each holding an estimated trajectory obtained by doubly integrating the accelerometer output. Boundary-matched splines were then fitted in each window to frame positions and accelerometer readings. Nützi *et al.* [15] fused IMU measurements into the tracking thread of Klein and Murray's PTAM [16]. Achtelik *et al.* [17] combined vision, IMU and pressure (hence height) measurements in an EKF to control a micro air vehicle, while Hide *et al.* [18] used IMU data with up-to-scale relative pose estimates from a downward facing camera for pedestrian navigation.

In the second category, a common assumption in vehicle navigation using visual odometry is that the camera is at a fixed and known height above the ground [19]–[22]. By detecting the ground plane, the local scale of the map is corrected to ensure that the camera height stays at its known value. Care must be taken determining the plane: [23] for example finds it by tracking objects likely to be resting upon it. As seen in later experiments, these constrained methods excel when this assumption holds, but are incapable of operating otherwise.

Another environmental assumption is that the scene contains discrete repeated structure. Botterill *et al.* [24] record descriptors of scene structure using the distances between characteristic features in it as measured from the image. When a similar pattern of features appears again, scale drift can be corrected. No prior knowledge of actual size is used, so that absolute scale is not recovered.

Strasdat *et al.* [25] describe a localisation and mapping method that is scale-drift aware. While pose-graph optimisation [26], [27] has a much lower computational complexity than full bundle adjustment, it is unable to correct scale drift on loop closure. Strasdat's method is an effort to bridge this gap by imposing similarity rather than rigid-body constraints between pairs of keyframes. When a loop is detected, drift is measured by comparing overlapping structure. This information is fed into the scale-factor of the loop-closure constraint which is propagated to the rest of the loop after optimisation. While the method shows improvement over rigid body pose-graph optimisation, the question of whether the method is better than a full bundle adjustment is left unanswered. More importantly, scale is only corrected *at* loop-closure, rendering it unsuitable for open-loop camera trajectories.

Castle *et al.* [28] incorporated planar objects identified by sets of SIFT features [29] into a map estimated using monoSLAM [30], [31]. Objects were detected live, a homography between the plane and the camera estimated and decomposed, allowing the object's position to be incorporated by monoSLAM's EKF. As the size of each object instance was known, the method could resolve the depth/speed scaling ambiguity. The requirement to see specific objects makes broader application problematic, but a further difficulty was that despite being able to set a global scale from object measurements, correcting scale-*drift* proved awkward in an EKF. Object measurements incorporated into the map at later times were found increasingly at odds with the scale of the surrounding landmarks, resulting either in tracking failure or in rejection of object detections. In later work [32] they used a bundle adjustment rather than an EKF, but objects were merely used as augmentations rather than for scale correction. Civera *et al.* [33] also used an EKF and augmented their map with full three-dimensional objects; but they too considered particular objects rather than object classes.

Three further works that make use of objects to improve maps are [34], [35], and [36]. Bao *et al.* [34] jointly estimate camera, point and rectangular object positions in a bundle adjustment. Their aim is to increase accuracy in batch-processed structure from motion. Fioraio and Di Stefano [35] have a similar aim, but adopt an incremental SLAM framework (albeit in a small environment and using RGBD imagery) and propose low-level matching to a database of characteristic features from objects rather than using an object detector. Neither [34] nor [35] is concerned with correcting scale over long term SLAM trajectories. More recently, Gálvez-López *et al.* [36] proposed adding objects modelled from point clouds to the map in which known distances between object-points are used for adding additional geometrical constraints to a bundle adjustment and enforcing scale in the map. As the method uses a library of specific object instances, its applicability to a more general setting is uncertain. Although the global scale of the map is metrically accurate, no results for long-term drift-free operation were shown in [36]. In the next section we pursue the idea of using pairwise constraints between point landmarks, but find that they are not effective for avoiding scale-drift in a real-time system.

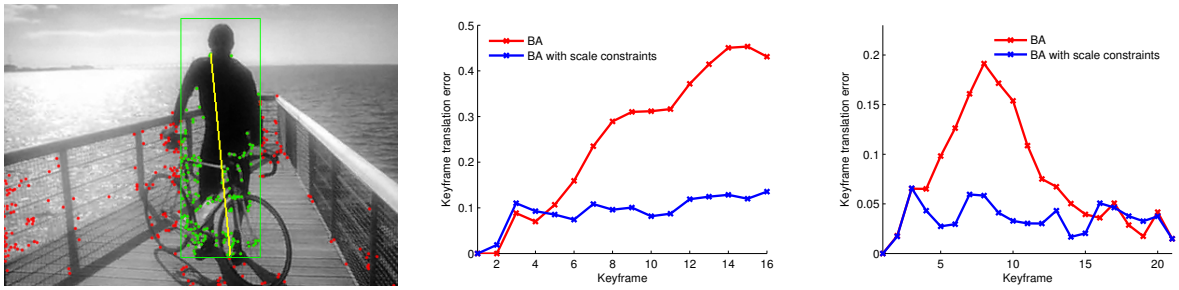
Dame *et al.* [37] used a monocular method to produce a dense map that is later refined using 3D shape priors embedded in a low-dimensional latent space. As the scale of shapes is known, the scale of the map may be set. Like [36] however, the method was tested in only a small localised map rather than with long range data, and it is unclear whether objects would be able to correct drift rather than just setting a global scale.

In this paper we draw on methods which supplement the map with objects, in preference to imposing some form of constraint on the sensor position or using additional sensing hardware. To avoid relying on viewing specific object instances, we will use object classes chosen both for their ubiquity and for low variance in their actual size. Their size distributions will be fixed beforehand to promote long-term accuracy. We also propose to use minimal object representation that is compatible with online scene and camera localisation using bundle adjustment.

III. L2L CONSTRAINTS: A FAILED REPRESENTATION

Accepting that object size can regulate scale does not answer the question of how best to incorporate size information into a bundle adjustment. Following [36], we first test the proposal that objects provide a metrically accurate coordinate system on which expected distances *between* landmarks may be computed.

Consider a set of landmarks $\mathcal{X}=\{\mathbf{X}_{iw}\}$, $i = 1, \ldots, I$ in the world frame observed in a set of camera keyframes with



(a) (b) (c)

Fig. 1. (a) Object detection for L2L constraint calculations. The bounding box sets up a 2D coordinate frame from which the distance between a pair of landmarks (highlighted in yellow) can be estimated. Per-keyframe translation errors compared for unconstrained (red) and L2L-constrained (blue) bundle-adjustments, for (b) an open trajectory and (c) a closed loop trajectory.

poses $\mathcal{T}=\{\mathtt{T}_k\}$, $k=1,\ldots,K$. Vanilla bundle adjustment seeks to find landmark positions and camera poses that minimise the total reprojection cost, the summed weighted 2-norms of the residuals $\tilde{\mathbf{x}}_{ik}=(\mathbf{x}_{ik}-\hat{\mathbf{x}}(\mathbf{X}_{iw},\mathtt{T}_k))$ between image measurements and predicted projections of landmarks i in cameras k. Assuming errors are normally distributed with zero mean and covariance $\mathtt{X}_{ik}$, bundle adjustment provides the maximum likelihood estimator of $\mathcal{X}$ and $\mathcal{T}$:

$$\begin{aligned}\{\mathcal{X},\mathcal{T}\}^* &= \underset{\{\mathcal{X},\mathcal{T}\}}{\arg\min}\, E_{\text{reproj}} \\ &= \underset{\{\mathcal{X},\mathcal{T}\}}{\arg\min} \sum_{i\in\mathcal{X}}\sum_{k\in\mathcal{T}} \tilde{\mathbf{x}}_{ik}^\top \mathtt{X}_{ik}^{-1}\tilde{\mathbf{x}}_{ik}\ . \end{aligned} \tag{1}$$

Our use of landmark-to-landmark (L2L) constraints requires the distance $d_{ii'}$ between landmarks i and i' to be drawn from some normally distributed prior with known mean $\mu_{ii'}$ and variance $\sigma^2_{ii'}$. Summing over all such available constraints gives a cost

$$E_{\text{constraint}} = \sum_{i,i'\in\mathcal{X}} \frac{1}{\sigma^2_{ii'}}(d_{ii'}-\mu_{ii'})^2 \tag{2}$$

which can be used to regularise the adjustment, as

$$\{\mathcal{X},\mathcal{T}\}^* = \underset{\{\mathcal{X},\mathcal{T}\}}{\arg\min}\,[E_{\text{reproj}} + \lambda E_{\text{constraint}}\,]\,. \tag{3}$$

Obtaining an L2L constraint from imagery involves detecting an instance of a known object class in keyframes. An example using humans is shown in Fig. 1(a). A part-based object detector (*e.g.* [38]) delivers a bounding box round the detection and, from pre-computed distributions on the object's width and height, and assuming planarity, distance constraints between landmarks within the object are obtained by 2D projection. In Fig. 1(a) for example, pre-tabulated mean dimensions of humans allows the distance between two landmarks to be calculated as $1.56\,\text{m}$, a reasonable value for that between shoulder and foot.

The approach was first tested in simulation with known ground truth. At the end of a trajectory (in which one constraint was made available per keyframe) both unconstrained and L2L-constrained bundle adjustments were performed. Figs. 1(b) and (c) compare the resulting errors in camera translation per keyframe for an open trajectory and a closed-loop trajectory, respectively. The constraints evidently reduce the per-keyframe error, keeping it broadly constant across the trajectory.

Despite this promise, when the L2L regulariser was tested live in a modified PTAM system [16] two difficulties became apparent. First, the planar assumption needed to obtain the constraints was too restrictive. Secondly, and more importantly, it was found that constraints introduced some time into the processing — and after the scale had drifted — were quite unable to restore the scale. Landmarks linked by constraints would move rapidly to obey them, but the consequential large changes in their position often resulted in bundle adjustment ignoring their measurements as outliers rather than re-scaling the map.

These difficulties resonate with those reported by Castle *et al.* [28] when imposing inter-landmark constraints into an EKF SLAM. Both outcomes indicate that sudden interventions at low-level are too disruptive of the probabilistic basis guiding the optimisation, whether recursively via a Kalman filter or in batch mode via bundle adjustment. Both suggest avoiding transference of somewhat coarse-scale information about object size directly onto fine-scale point landmarks already in the map. A method which achieves this is developed next.

IV. Bundle Adjustment Using Object Landmarks

Section II reviewed a number of object parameterisations that have been introduced into SLAM, ranging from the bounding boxes of [39] to the 3D surface models in [35], [37]. While more complicated models certainly allow for accurate segmentation, and can improve the quality of reconstruction locally, they impose a considerable computational overhead. Even rectangular bounding boxes require a full 6 degree-of-freedom (dof) pose to be maintained, while 3D surface models require 2D silhouette segmentations to be localised correctly. While the results of such reconstructions are visually impressive, we suggest that their additional complexity is unnecessary if the sole aim is scale correction. Instead, here we propose representing both objects and points as generalised landmarks with different "extents".

A. Object representation

The combined representation of object and point landmarks in the scene and image is shown in Fig. 2. Both are represented

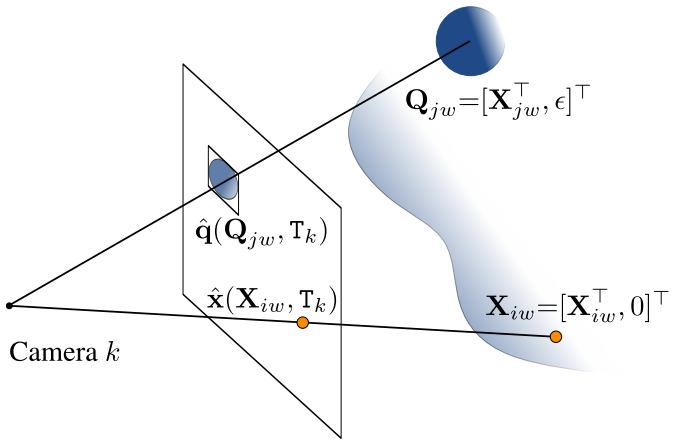


Fig. 2. Parameterisations of points and objects with their respective projections.

by a minimum enclosing sphere, and have a single additional size parameter, the sphere's radius. We shall refer to this as the extent ϵ. An object j is thus represented in the world frame by

$$\mathbf{Q}_{jw} = \begin{bmatrix} \mathbf{X}_{jw} \\ \epsilon_j \end{bmatrix} \tag{4}$$

where $\mathbf{X}_{jw} = [X, Y, Z, 1]_{jw}^\top$ is the homogeneous location of its centre. As the object is deemed spherical, transforming between coordinate frames is simple: the extent is invariant. An object with coordinates $\mathbf{Q}_{jw}$ in the world frame becomes

$$\mathbf{Q}_{jk} = \begin{bmatrix} \mathbf{X}_{jk} \\ \epsilon_j \end{bmatrix} = \begin{bmatrix} \mathtt{T}_k & \mathbf{0}^{4\times1} \\ \mathbf{0}^{1\times4} & 1 \end{bmatrix} \mathbf{Q}_{jw} \tag{5}$$

in camera frame k. The projection of the object into this image is the 4-vector

$$\hat{\mathbf{q}}_{jk} = \mathtt{K}\,\mathrm{proj}(\mathbf{Q}_{jw}, \mathtt{T}_k) = [u, v, w, h]_{jk}^\top \ , \tag{6}$$

where $[u, v]$ are the image coordinates of the projection of the centre of the object and $\mathtt{K}$ is the camera's intrinsic matrix, with focal lengths f_u and f_v and principal point $[u_0, v_0]$. The parameters $[w, h]$ are the width and height of the projected bounding box in the image, and are given by

$$\begin{bmatrix} w \\ h \end{bmatrix}_{jk} = 2\epsilon_j Z_{jk}^{-1} \begin{bmatrix} f_u \\ f_v \end{bmatrix} \ . \tag{7}$$

B. Object measurements and data association

Objects are localised using a detector applied to each new keyframe as it is created. Their measurements are found from the bounding box around the detection and parameterised in the same way as object reprojections by the 4-vector $\mathbf{q}_{jk} = [u, v, w, h]_{jk}^\top$. For solving data association between keyframes we use the method described in [40] and [41].

C. Object Bundle Adjustment

If object measurements are distributed with covariance $\mathtt{Q}_{jk}$ a new bundle adjustment that seeks to find the most likely sets of objects $\mathcal{Q} = \{\mathbf{Q}_{1w}..., \mathbf{Q}_{Kw}\}$, point landmarks $\mathcal{X}$ and keyframes $\mathcal{T}$ may be written as

$$\{\mathcal{X}, \mathcal{Q}, \mathcal{T}\}^* = \underset{\{\mathcal{X},\mathcal{Q},\mathcal{T}\}}{\arg\min} \Bigg(\sum_{i\in\mathcal{X}} \sum_{k\in\mathcal{T}} \tilde{\mathbf{x}}_{ik}^\top \mathtt{X}_{ik}^{-1} \tilde{\mathbf{x}}_{ik} + \sum_{j\in\mathcal{Q}} \sum_{k\in\mathcal{T}} \tilde{\mathbf{q}}_{jk}^\top \mathtt{Q}_{jk}^{-1} \tilde{\mathbf{q}}_{jk} \Bigg) \ , \tag{8}$$

where $\tilde{\mathbf{q}}_{jk} = (\mathbf{q}_{jk} - \hat{\mathbf{q}}(\mathbf{Q}_{jw}, \mathtt{T}_k))$.

However, the adjustment can be written more economically. We first assume a known extent for a particular object class, and fix the extent of each object instance of the class to this value. Only the object's position in the 3D map requires refinement in normal operation. Consequently, as point landmarks have known zero extent, $\mathcal{X}$ can be treated as a subset of $\mathcal{Q}$.

The noise covariances for point landmarks and objects are also unified. Experiment indicates that we can take the extent to be normally distributed as $\mathcal{N}(\epsilon, \sigma_\epsilon^2)$. Using Eq. (7) the projected extent in frame k is also approximately normally distributed as

$$\begin{bmatrix} \hat{w} \\ \hat{h} \end{bmatrix} \sim \mathcal{N}\left(\epsilon\mathbf{f}, \sigma_\epsilon^2 \mathbf{f}\mathbf{f}^\top\right) \ , \tag{9}$$

where

$$\mathbf{f} = 2\langle Z_{jk}^{-1} \rangle \begin{bmatrix} f_u \\ f_v \end{bmatrix} \ . \tag{10}$$

Writing the measurement process on width and height as $\sim\mathcal{N}(\mathbf{0}, \Sigma_{\mathrm{box}})$, then the measured width and height will be distributed as

$$\begin{bmatrix} w \\ h \end{bmatrix} \sim \mathcal{N}\left(\epsilon\mathbf{f}, [\sigma_\epsilon^2 \mathbf{f}\mathbf{f}^\top + \Sigma_{\mathrm{box}}]\right) \ . \tag{11}$$

Assuming independent zero-mean, noise on the centre of the bounding box with variances $\sigma_{x,y}^2$, the covariance for object and point landmark measurements becomes

$$\mathtt{Q}_{jk} = \begin{bmatrix} \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix} & \mathbf{0}^{2\times2} \\ \mathbf{0}^{2\times2} & \sigma_\epsilon^2 \mathbf{f}\mathbf{f}^\top + \Sigma_{\mathrm{box}} \end{bmatrix} \ , \tag{12}$$

where the bottom right 2×2 matrix defaults to zero for point landmarks. Numerical values are considered later.

The adjustment in Eq. (8) then becomes just

$$\{\mathcal{Q}, \mathcal{T}\}^* = \underset{\{\mathcal{Q},\mathcal{T}\}}{\arg\min} \sum_{j\in\mathcal{Q}} \sum_{k\in\mathcal{T}} \tilde{\mathbf{q}}_{jk}^\top \mathtt{Q}_{jk}^{-1} \tilde{\mathbf{q}}_{jk} \ . \tag{13}$$

This simplification allows for a Hessian structure that is identical to general bundle adjustment. Although there are additional operations to deal with the extra parameters in the measurement residuals, the main computational cost comes from solution of the normal equations, which is no more than bundle adjustment using landmarks alone.

V. Implementation

The combined representation of point and object landmarks can be applied to any sparse keyframe-based SLAM system. We discuss implementational details which have been found essential for long term operation.

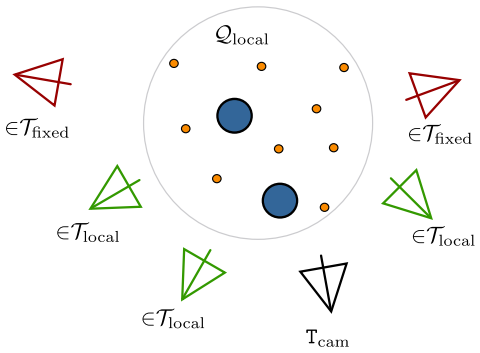


Fig. 3. Local bundle adjustment centred on $\mathtt{T}_{\text{cam}}$. Points and object landmarks measured in $\mathtt{T}_{\text{cam}}$ and up to 10 most proximate keyframes are included, as are any other keyframes with measurements of those landmarks. These latter keyframes are held fixed. The number of object landmarks used in practice is a tiny fraction of the number of regular scene points.

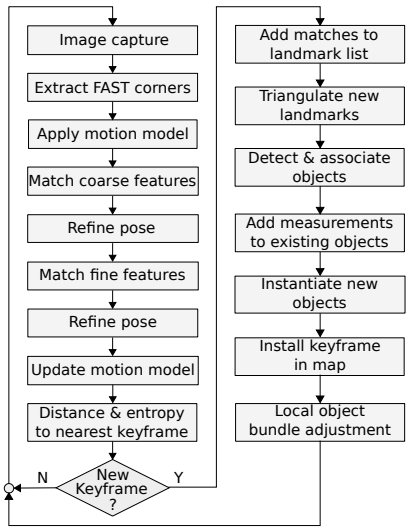


Fig. 4. The processing pipeline. The framerate tracking pipeline is in the left column, and the processes of adding a new keyframe ending with object bundle adjustment is in the right column.

A. Local adjustment

Applying a global bundle adjustment to every frame in a video sequence quickly becomes computationally impracticable. Instead, as illustrated in Fig. 3, we optimise a local area around the current estimated camera position, giving a broadly constant computational load consistent with video rate operation. A local set $\mathcal{T}_{\text{local}}$ of n nearest keyframes (we use $n{=}10$) is defined around the current camera $\mathtt{T}_{\text{cam}}$. The structure to be refined consists of the set $\mathcal{Q}_{\text{local}}$ of landmarks and objects visible to the current camera *and* the local set. To constrain the local adjustment within the rest of the map, all observations of $\mathcal{Q}_{\text{local}}$ from other keyframes are included in the optimisation. These other keyframes form a set $\mathcal{T}_{\text{fixed}}$ whose poses are held fixed during optimisation. The local optimisation is

$$\{\mathcal{Q}_{\text{local}}, \mathcal{T}_{\text{local}}, \mathtt{T}_{\text{cam}}\}^* = \underset{\{\mathcal{Q}_{\text{local}}, \mathcal{T}_{\text{local}}, \mathtt{T}_{\text{cam}}\}}{\arg\min} \sum_j \sum_k \tilde{\mathbf{q}}_{jk}^\top \mathtt{Q}_{jk}^{-1} \tilde{\mathbf{q}}_{jk} \ , \tag{14}$$

where $j{\in}\mathcal{Q}_{\text{local}}$ and $k{\in}\{\mathtt{T}_{\text{cam}}, \mathcal{T}_{\text{fixed}}, \mathcal{T}_{\text{local}}\}$. Because the adjustment is run in just a local area around the current camera position, it is unable to propagate corrections from the local window to the rest of the map. If scale drift occurs due to a lack of object detections over a period, only the most recent portion of the map consisting of the n keyframes in the local window will be corrected when objects are seen again.

B. Noise reduction

Actual object measurements are subject to a number of sources of error: principally, false positives in object detection; inaccurate size and location in the image; incorrect association between otherwise correct detections; and, most importantly, movement independent of the sensor between frames. The following are used to mitigate their effects.

First, unlike point landmarks, which are essential for camera tracking and must be localised as soon as possible, object measurements need not be used promptly. Instead, we accumulate a minimum number of measurements for each object before using them in the adjustment. Several benefits accrue: (i) the increased number of measurements ensures only well-localised objects are used; (ii) any false positives from the object detection algorithm will be ignored provided they fall under this threshold; and (iii), unless they are in front of the camera and moving at the same speed, moving objects will have less than the required number of measurements and will consequently be ignored.

Second, to down-weight poor object measurements or poor data association, object and point landmark measurements are wrapped in a Tukey biweight objective function (*c.f.* [42]). Eq. (13) becomes

$$\{\mathcal{Q}, \mathcal{T}\}^* = \underset{\{\mathcal{Q}, \mathcal{T}\}}{\arg\min} \sum_{j\in\mathcal{Q}} \sum_{k\in\mathcal{T}} \text{Obj}\left(|\tilde{\mathbf{q}}_{jk}|, \mathtt{Q}_{jk}, \sigma_T\right) \ . \tag{15}$$

We consider three cases. For point landmark measurements, σ_T is set equal to the estimated standard deviation of the distribution of point errors. For the bounding box position $(x, y)_{jk}$ in measurements $\tilde{\mathbf{q}}_{jk}$, σ_T is set to accommodate its inherent lower accuracy compounded with the possibility that scale drift is as yet untamed. Experiment suggests a value at least an order of magnitude higher than the standard deviation found in the error distribution after scale has stabilised (Fig. 12(a)). For the bounding box size, $(w, h)_{jk}$, we find it best to turn off robust weighting. If objects have been observed recently the measurement values are well-behaved statistically (see Fig. 12(b)), but if not, scale drift may cause objects to appear disproportionately large or small in the surrounding map. For drift correction it is imperative that the adjustment not reject these measurements.

Note that an object moving at the same speed and direction as the camera incorrectly implies a stationary camera. However, when there are sufficient landmark measurements that disagree with this inference, and the camera is indeed moving, the object will be considered an outlier by the robust estimator.

C. Embedding in a tracking and mapping method

At the system level, illustrated in Fig. 4, we apply the method by modifying a basic version of PTAM [16]. After

initialisation, the camera pose tracker runs continuously every frame, using a simple motion prediction to assist matching of FAST features [43] to landmark projections and thence iteratively to optimise the pose using a re-weighted least squares algorithm. Like PTAM, the method matches features using 8×8 pixel templates that are first coarsely matched with patches from other features in a search radius and then iteratively refined for sub-pixel precision.

To determine whether a keyframe is required, in addition to measuring the distance from other keyframes we monitor the entropy ratio between the current camera estimate and the most recent keyframe, as suggested by Kerl *et al.* [2]. This measure, unlike the distance metric, has the benefit of being invariant to scale drift. If either metric crosses a threshold, the current frame is tagged as a keyframe.

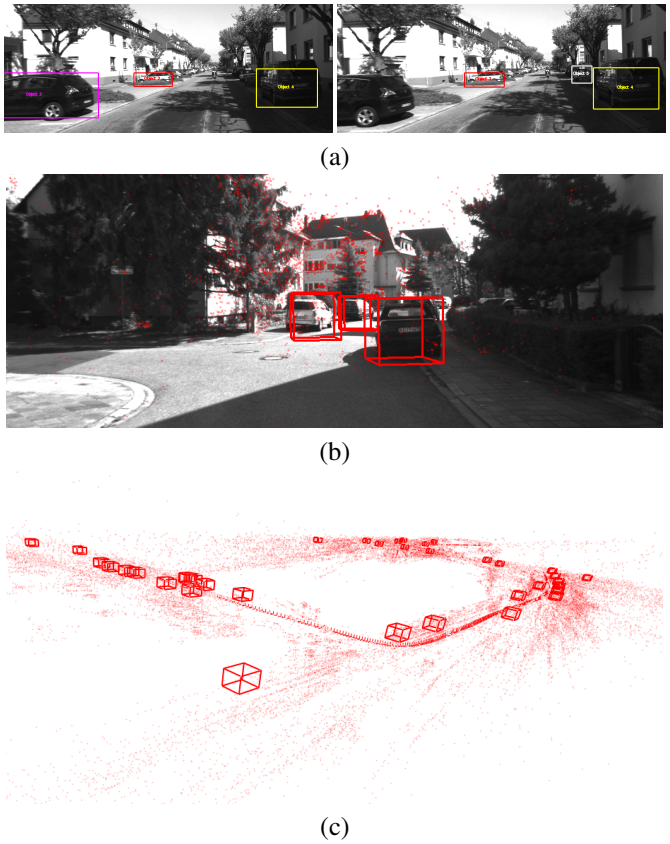
As only a subset of landmarks is used for tracking, the mapmaker first searches for measurements of other landmarks and adds them to the keyframe. A set of epipolar matches is found with the closest neighbouring keyframe using PTAM's patch matching algorithm, and new 3D map points triangulated. Any object measurements in the form of bounding boxes from object detections are then added to the keyframe. If an object has not been seen before, it can be localised immediately from a single measurement in the keyframe (as its projection is invertible). However it is not allowed into the bundle adjustment until its number of measurements exceed the threshold discussed in Section V-B.

If there are enough object measurements in the surrounding keyframes, a local landmark and object bundle adjustment is applied. If no object is found, or the currently visible objects have insufficient measurements, a local point landmark-only bundle adjustment is performed. With a window of 10 adjustable keyframes the adjustment takes around 150 ms to converge and so tracking is not disrupted. The new keyframe is then added to a queue in the mapmaker, which adds it to the map in a separate thread. Finally the motion model is updated with the current position of the camera.

VI. Experimental results

The performance of the method has been evaluated on kilometre-long outdoor sequences from the KITTI street scene dataset [20], [44], and on 100-metre long outdoor sequences from hand-held phone cameras. An illustration of the overall performance on one such sequence is shown in Fig. 5, where (a) shows object detections, panel (b) shows a keyframe with both point landmark and object detections and (c) shows the recovered map. It will be evident from Fig. 5(c), and later from Fig. 14(c), that the number of object landmarks is a small fraction of the number of point landmarks. A typical value is 0.1%. Improvements in the overall structure result from objects stabilising scale, rather than their adding to the mere bulk of measurements.

Our method is able to deal with multiple arbitrary object classes, indeed any for which an extent distribution is known. However, the nature of the KITTI dataset makes cars the obvious choice here. Object detections in keyframes and corresponding data association labels are obtained using the



(a) (b) (c)

Fig. 5. Examples of (a) object detections from successive frames; (b) objects and point landmarks used in tracking; and (c) a 3D map with objects and points generated by the method, all from a sequence from the KITTI dataset.

tracking-by-detection algorithm of Zhang *et al.* and Geiger *et al.* [40], [41], a method tailored to the detection of vehicles in street settings. The extent for cars was set at $\bar{\epsilon}$=1.2 m, an average over popular European makes.

To explore different aspects of the method's performance we report the outcome of individual experiments as follows:

- *A.* We compare the performance of object bundle adjustment with one that ignores objects and has no scale information.
- *B.* The performance of the proposed method is evaluated when there are few objects in part of the sequence.
- *C.* We examine the relationship between number of object observations and the error in camera speed.
- *D.* The size distribution and noise model is validated by using ground truth camera pose data from the KITTI dataset.
- *E.* An online tool is used to compare the performance of our method compared with those monocular odometry methods which perform best on KITTI data.
- *F.* We present results from a video sequence captured with a hand-held camera moved in an unconstrained way.

A. Output with and without objects

To demonstrate the scale-correcting nature of object-supplemented bundle adjustment, the algorithm is run on a number of video sequences, first using landmark-only bundle adjustment and then using object-supplemented adjustment.

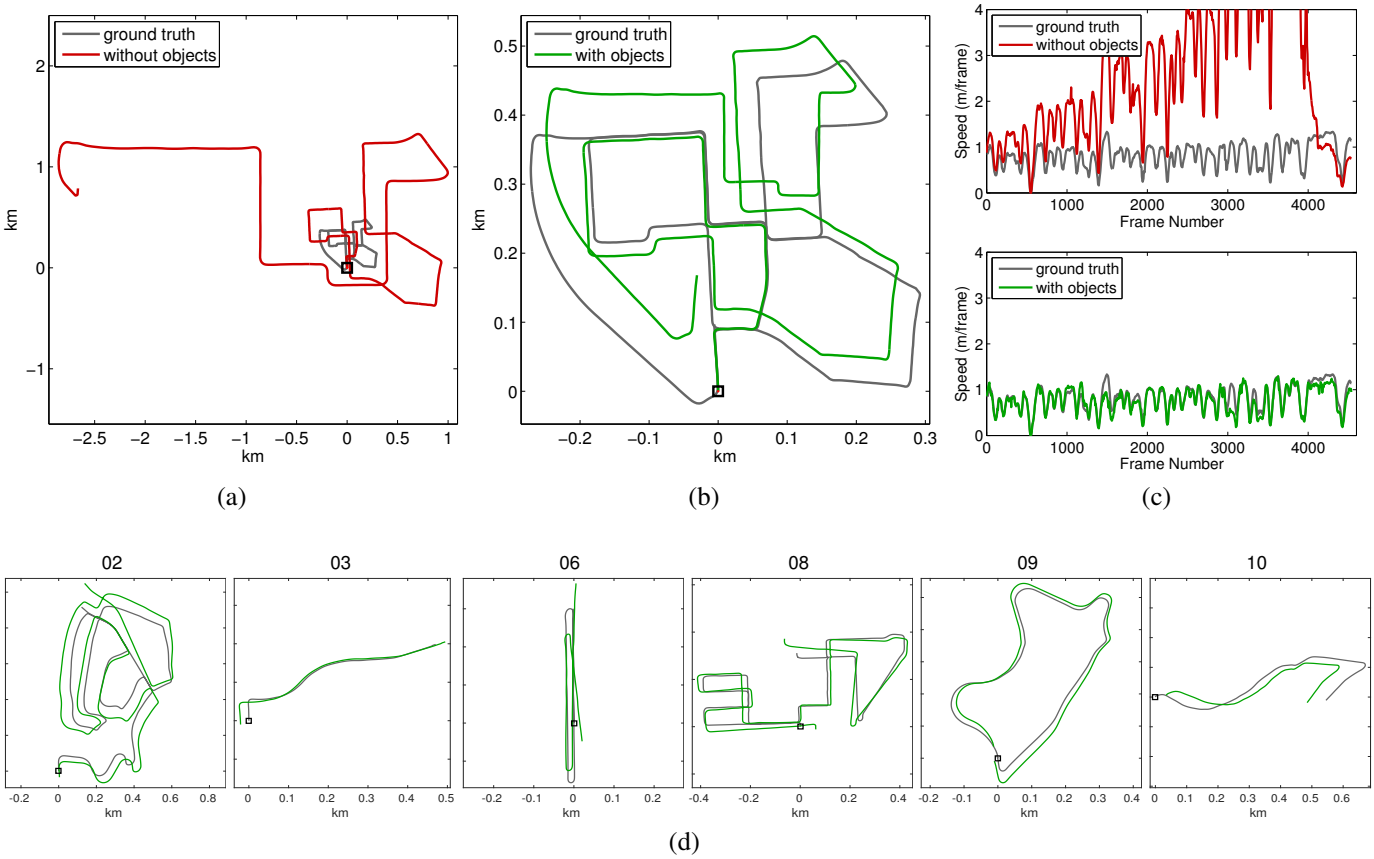


Fig. 6. Comparisons between the ground truth trajectory (grey) for KITTI sequence 0 and those obtained using (a) bundle adjustment with points alone and (b) bundle adjustment supplemented with objects. The speeds per frame (c) without and with objects compared with ground truth. Trajectories (d) from running object-supplemented bundle adjustment on other KITTI sequences.

Fig. 6 shows a comparison of the resultant trajectories obtained using (a) point landmark-only bundle adjustment with (b) our proposed point and object bundle adjustment, both applied to KITTI sequence #0. While the point landmark adjustment accumulates a massive amount of scale drift over the course of the trajectory, the object-supplemented version is able to maintain a map that is within essentially the same scale as that of ground truth — this without taking advantage of the several opportunities for loop closure. The reduction in error is more clearly seen in Fig. 6(c) which shows the speed of the camera compared to ground-truth for both methods. The camera speed at keyframe k was calculated as a backwards difference

$$s_k = |\mathbf{c}_k - \mathbf{c}_{k-1}| \ , \tag{16}$$

where $\mathbf{c} = \mathtt{R}^\top \mathbf{t}$ is the keyframe's camera centre in world co-ordinates. The difference in estimated and true camera speeds grows over time with the landmark-only bundle adjustment, while the object-supplemented version stays close to ground truth. We stress that although there are loops in the real camera trajectory, we intentionally do not close them, so as to generate longer open-loop trajectories. Fig. 6(d) shows the resultant trajectories from running the object-bundle adjustment on further KITTI sequences.

Table I compares a root mean square error E_{RMS} for the trajectories recovered with and without objects for the first 11 sequences in the KITTI dataset which come with ground truth

TABLE I
RMS ERROR IN TRANSLATION OVER A TRAJECTORY WITHOUT OBJECTS AND WITH OBJECTS (THIS WORK) FOR THE SEQUENCES 0-10 IN THE KITTI DATASET.

Sequence number	Number of frames used	E_{RMS} without objects	E_{RMS} with objects
0	4541	1181.0	**73.4**
1	1101	712.0	**545.8**
2	4661	815.7	**55.5**
3	801	81.1	**30.6**
4	271	**7.4**	10.7
5	2761	798.8	**50.8**
6	1101	244.9	**73.1**
7	1101	110.2	**47.1**
8	4071	1907.6	**72.2**
9	1591	139.6	**31.2**
10	1201	115.3	**53.5**

data. E_{RMS} is a scale-invariant measure of the translation error for a monocular sequence, given by (*c.f.* [25])

$$E_{\text{RMS}} = \left[\frac{1}{N} \arg\min_s \sum_{k \in \mathcal{T}} (\mathbf{t}_k - s\hat{\mathbf{t}}_k)^2 \right]^{1/2} \ , \tag{17}$$

where s is a global scale factor applied to the estimated trajectory, and $\mathbf{t}_k$ is the translational component of keyframe k. Because ground truth is available, Table I shows values with s=1, but with the estimated trajectory already normalised such that its *starting* scale is equal to that of the ground-truth. For

Fig. 7. Example frames from KITTI sequences #1 and #4 which are problematic as they contain few stationary vehicles.

all but two sequences, the use of objects results in a marked decreased in translation error. Sequences #1 and #4 continue to have a high amount of error: there is a lack of stationary objects in these sequences as illustrated in Fig. 7.

B. Recovery after scale drift

The principal failure of the method of Section III was that the introduction of scale correction *after* substantial drift had occurred (because of a lack of object measurements, say) was disruptive, causing good measurements to be rejected. Fig. 8 demonstrates that this is not the case now. It shows the estimated trajectory and camera speed obtained by ignoring all object measurements for the first 2000 frames. Almost immediately after objects are used again the scale is corrected to a value close to ground truth.

C. Effect of object observations on speed estimation

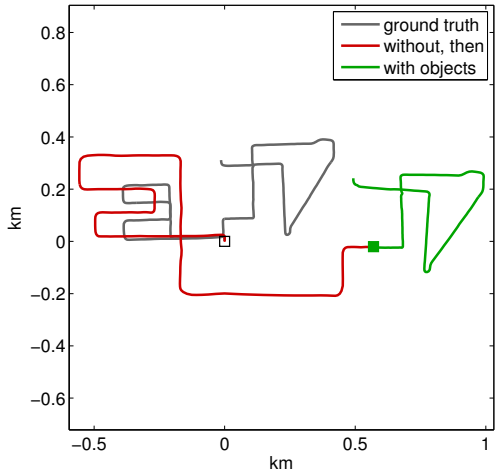
While Section VI-A shows that there is a clear advantage of using object-supplemented bundle adjustment over its point landmark-only counterpart, here we present a more localised view of how objects affect scale estimation.

For each keyframe in a trajectory, the RMS error in speed is computed along with the average number of object observations that are present in a 10-keyframe neighbourhood. The averaging is over space, as observations typically affect a number of surrounding keyframes rather than a single one. A single keyframe with no object observations but surrounded by keyframes with observations is still likely to have a low speed error, despite having no observations itself.

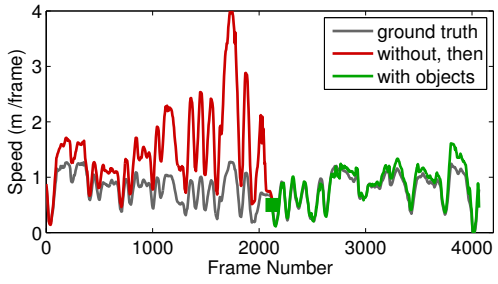
Fig. 9 shows a plot of the maximum error in the speed versus the average number of observations for KITTI sequences 0-10, but excluding sequences #1 and #4 because of their lack of stationary objects. Increasing the number of observations does indeed decrease the error's upper bound.

D. Validating the object extent distribution

As noted earlier, it is not our aim to learn the size or structure of the detected objects at run-time, but to use pre-tabulated size data to stabilise scale. In this work we have used manufacturers' data to find a mean extent $\bar{\epsilon}$ of popular makes of vehicles — but size data in other objects may be found in sources such as architectural handbooks, (*e.g.* [45]), the human studies literature (*e.g.* [46]), and so on. However, because the KITTI dataset provides known camera poses at each frame,



(a)



(b)

Fig. 8. Estimated trajectory (a) and translational speed (b) obtained by initially ignoring objects in the first 2100 frames of the KITTI sequence #8. After this point, marked in green, objects are included in the bundle adjustment, and scale is promptly corrected.

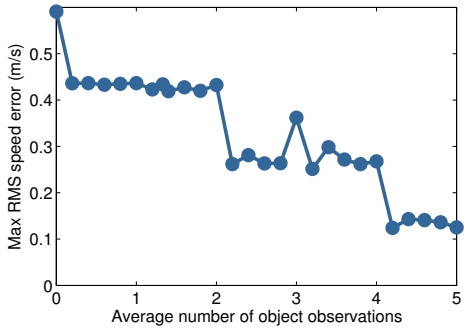


Fig. 9. Plot of the maximum RMS speed error versus the number of object observations per neighbourhood for KITTI sequences that contain stationary cars.

we are able to validate the proposed extent distribution for cars in several ways.

1) Using the extent as discrete parameter: First we consider an empirical search over ϵ, treating it as a parameter to be optimised by evaluating the performance of the system on all sequences of the KITTI dataset that contain cars. The performance metric chosen is the average per-frame RMS error in estimated speed

$$E_{\text{speed}}(\epsilon) = \left[\frac{1}{N}\sum_{i=1}^{N}\left(s_i - \hat{s}_i(\epsilon)\right)^2\right]^{1/2}, \qquad (18)$$

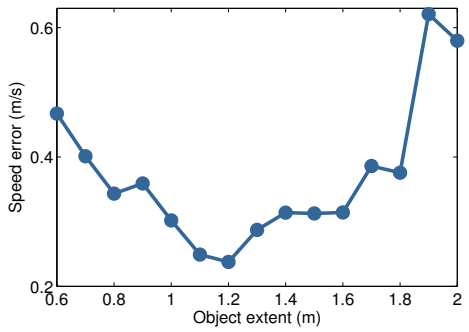


Fig. 10. The averaged per-frame RMS error in speed plotted against object extent ϵ derived from all KITTI sequences that contain cars.

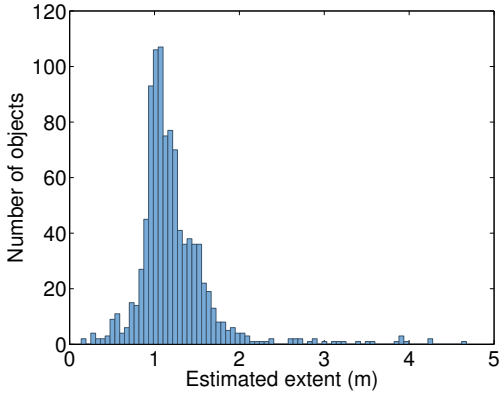


Fig. 11. Frequency of estimated object extent in the KITTI dataset.

where s_i is the ground-truth speed from Eq. (16), $\hat{s}_i(\epsilon)$ is our estimated speed as a function of extent, and N is the total number of keyframes used.

Fig. 10 shows the error for various object extent values. The value with the lowest error occurs at $\epsilon = 1.2\,\mathrm{m}$, in agreement with the mean value from manufacturers' data.

2) Recovering an extent distribution: A second validation estimates a distribution over ϵ directly, again exploiting the ground-truth poses from KITTI. Given a set of ground truth keyframe poses $\mathcal{T}_{GT}$ with associated object observations, a set of objects $\mathcal{Q}^*$ from the particular class in question is estimated by minimising

$$\mathcal{Q}^* = \arg\min_{\mathcal{Q}} \sum_{j \in \mathcal{Q}} \sum_{k \in \mathcal{T}_{GT}} \tilde{\mathbf{q}}_{jk}^\top \tilde{\mathbf{q}}_{jk} \ . \tag{19}$$

This is simply the object term of Eq. (8) without an associated covariance matrix. The extent ϵ of each object is no longer a constant for the entire class, but is a free parameter estimated, alongside location, for each object instance.

Fig. 11 shows the resultant distribution for cars detected in KITTI sequences #0 to #10. The distribution has mean $1.2\mathrm{m}$ and variance $0.2\mathrm{m}^2$, in agreement with earlier estimates. Moreover, the assumption of a normal distribution appears justified.

3) The errors in object box dimensions: The distributions for object-related errors (Eq. 12) are generated by accumulating histograms of the differences between the position and size of the bounding box as detected in each image and the position and size given by bundle adjustment after convergence, and taking the latter as correct.

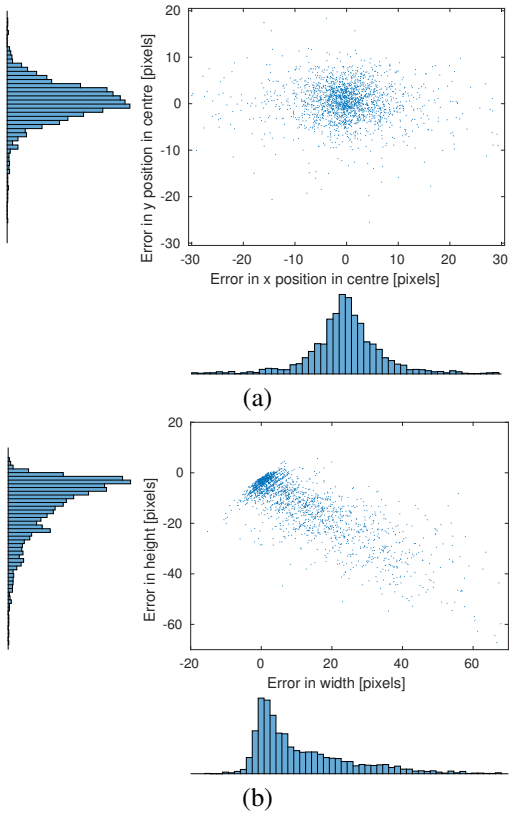


Fig. 12. Scatter plots of error in (a) location of object detections and (b) object bounding box width and height.

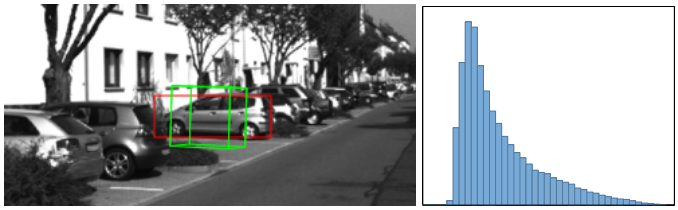


Fig. 13. (a) The fit (in green) to our spherical object model will tend to truncate the width and increase the height of a fully elongated detection (in red). (b) A synthesised distribution of width errors reproduces those observed in Fig. 12(b).

Histograms of errors in location (x, y) are shown in Fig. 12(a). They have zero-mean, are independent, and close to normal, thus adhering to the assumptions of Eq. (12). The standard deviations σ_x and σ_y are estimated as 6.6 pixel and 4.1 pixel, respectively. The histograms of errors in (w, h) are shown in Fig. 12(b). The means of the errors in width and height are 10.4 and -11.6 respectively. The error covariance is

$$\sigma_\epsilon^2 \mathbf{f}\mathbf{f}^\top + \Sigma_{\text{box}} = \begin{bmatrix} 190.0 & -123.4 \\ -123.4 & 128.2 \end{bmatrix} \ . \tag{20}$$

The signs of the means arise because in KITTI both the camera's x-axis and each car's major axis lie predominantly in the horizontal plane. Fitting to an object model with a single extent will tend to pinch in the width and push up the height,

Fig. 14. (a) Stills from a handheld video sequence with varying camera height: high above, low below. (b) Resultant trajectories estimated using point-only bundle adjustment without objects (red), and object-supplemented bundle adjustment (green). Satellite imagery (Google) of the mapped area is shown for comparison. (c) The point cloud recovered using objects and (d) on enlargement the detected vehicles are visible in green.

as illustrated in Fig. 13(a). The asymmetric distributions in Fig. 12(b), which somewhat break the normal assumption, arise because detections closer to the object (i) are subject to greater perspective distortion but (ii) are fewer in number. The distribution synthesised using realistic assumptions about the camera's optics, its speed, and the detection of objects is shown in Fig. 13(b) and reproduces that observed in practice in Fig. 12(b).

E. An online assessment

Our method has been evaluated using KITTI's online assessment tool[1] [44] and its reserved testing sequences. Unfortunately, all the reference methods there employ a planar constraint in one form or other, making the comparison less than fair on our unconstrained approach. The average performance over all constrained monocular SLAM systems is captured by some 9% translation error and 0.020 deg m^{-1} rotational error, while the best performing constrained method achieves 2.4% translation error and 0.006 deg m^{-1} rotational error [22], [23]. Our unconstrained method yields 20% translation error and 0.014 deg m^{-1} rotational error on average, values which at least approach, and for rotation exceed, the average of the constrained methods.

Although our method will struggle to compete with constrained methods in absolute terms, there remains scope for further reducing error. First, the large inter-frame motion in the KITTI sequences over-stretches the feature and camera tracking stage of PTAM which was developed specifically for small AR workspaces. A simple KLT-based [47] odometry system is able to find numerous matches between every pair of frames on sequences where PTAM lost a number of map points. Sequence #1 in Table I is one such example. Second, a number of KITTI's *testing* sequences do not contain many static cars, making drift correction difficult. An extreme example is test sequence #14, which is captured in a park: there are shrubs but no cars. We discuss the use of multiple classes in our conclusions.

[1]at http://www.cvlibs.net/datasets/kitti/eval_odometry.php

F. Sequence from a hand-held: varying camera height

We suggested that the comparison against the state of the art provided by the KITTI online assessment tool was not a fair test of the current method because the best performers all constrain their cameras to have fixed height above the roadway.

To demonstrate the ability of our method to function without such a constraint a sequence was captured outdoors using a hand-held phone camera (a Samsung Galaxy S6 with fixed focal length, calibrated using standard methods). As ground truth was not available, the sequence contained a small loop, and performance was evaluated on the qualitative distance between the start and end of the loop. Cars were once again used as the object class and $\bar{\epsilon}$=1.2 m used as their extent.

Two frames taken from different camera heights within the sequence are shown in Fig. 14(a). The resultant trajectories estimated using a landmark-only and an object-supplemented bundle adjustment, respectively, are shown in Fig. 14(b) overlaid onto satellite imagery of the area where the sequence was filmed. Object-supplemented bundle adjustment provides a marked decrease in scale drift over its landmark-only counterpart. A detailed point cloud of the mapped area is shown in Fig. 14(c). Note again that the ratio of object points to background scene points is very small.

VII. Conclusion and Discussion

This paper has presented a novel method of incorporating scale information from object classes into monocular visual

SLAM using bundle adjustment. While other methods rely on further sensing hardware or assumptions about the height of the camera above the ground plane, the proposed method assumes only that the camera can observe known objects, and that those objects are can be reasonably described visually by a single extent. By avoiding recovery of object orientation, the structure of the bundle adjustment for objects is unified with that for regular scene points, resulting in no increase in computational complexity and only modest increase in computational cost.

It has been shown that the method is able to maintain a consistent scale estimate throughout a trajectory, provided regular object observations are available. The number of object landmark observations required is a remarkably small fraction of the number of point landmarks — in our experiments a typical fraction is 0.1%. We have demonstrated that when scale drift occurs through lack of object observations, it is rapidly reduced when objects are re-introduced.

Several avenues of work remain open for further consideration. First is the use of multiple object classes. Although we have used a single object class in this work, using multiple object classes is straightforward: one runs suitable object detectors and trackers in parallel, and thereafter, but with one caveat, bundle adjustment is indifferent to the object type. The caveat is that it becomes important for the respective size *distributions* to be known, so that proper differential weight can be given to size information from the several sources. We think it most likely that object extents will be found from independent data, but in this paper have demonstrated how values can be validated using visual information when ground truth camera motion is available.

A second lies in the detail of the SLAM process. At present the method cannot correct scale in parts of the map outside of the local bundle adjustment window. In Section VI-B for example, the bundle adjustment is not able to correct parts of the trajectory where no object observations are available. Strasdat *et al.* [48] have shown that a global bundle adjustment is unnecessary as accurate structure is only required around the current camera estimate. Rather, they apply a joint double-window optimisation that refines structure and keyframes in a local adjustment window, and a pose-graph optimisation to the rest of the map. One might explore whether scale information from a local object-bundle adjustment can be propagated correctly along the pose-graph using similarity transformations [25] in this formulation. This would in addition allow for simple closure of loops.

A third avenue concerns object modelling. Our results indicate that using an model with a single extent is not a hindrance to overall scale recovery. This is in part a benign outcome of integration of views around an object — the thin end of a long and thin object, for example, is viewed relatively infrequently; and in part a result of the clustering of visual interest. The simplest, and we suspect the most effective, enhancement to modelling would be to use a part-specific detector and to allocate different extents to different parts — the rear of a car, the side of a car, and so on. Using more complicated, pose-sensitive shape models is not without difficulty. For example, an ellipsoidal hull model might be used built either from non-visual measurements, or as shown recently in [49] and [50] using structure from motion with bounding boxes. Then, however, the 6 dof pose of the object needs to be estimated rather than just its 3 dof position, a requirement which will not sit straightforwardly in the same bundle adjustment framework as points. Further, as seen in [50], the recovery of pose from a few views as the camera passes by is not always reliable.

ACKNOWLEDGEMENTS

This work was supported by grant EP/J014990 from the UK's Engineering and Physical Science Research Council and Huawei Technologies Co. Ltd.

REFERENCES

[1] S. Kohlbrecher, O. von Stryk, and J. Meyer, "A flexible and scalable SLAM system with full 3D motion estimation," pp. 155–160, 2011.

[2] C. Kerl, J. Sturm, and D. Cremers, "Dense visual SLAM for RGB-D cameras," in *Proc 26th IEEE/RSJ Conf on Intelligent Robots and Systems*, pp. 2100–2106, 2013.

[3] C. Mei, G. Sibley, M. Cummins, P. Newman, and I. Reid, "RSLAM: a system for large-scale mapping in constant time using stereo," *International Journal of Computer Vision*, vol. 94, no. 2, pp. 198–214, 2011.

[4] J.-S. Gutmann and K. Konolige, "Incremental mapping of large cyclic environments," in *Proc IEEE Int Symp on Computational Intelligence in Robotics and Automation*, pp. 318–325, 1999.

[5] A. J. Davison and D. W. Murray, "Sequential localisation and map-building using active vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 865–880, July 2002.

[6] S. Se, D. G. Lowe, and J. J. Little, "Local and global localization for mobile robots using visual landmarks," *Proc 15th IEEE/RSJ Conf on Intelligent Robots and Systems, Lausanne, Switzerland, Oct 2-4, 2002*, pp. 414–420, 2002.

[7] M. Bosse, P. Newman, J. Leonard, and S. Teller, "Simultaneous localization and map building in large-scale cyclic environments using the atlas framework," *International Journal of Robotics Research*, vol. 23, no. 12, pp. 1113–1139, 2004.

[8] K. L. Ho and P. Newman, "Loop closure detection in SLAM by combining visual and spatial appearance," *Robotics and Autonomous Systems*, vol. 54, pp. 740–749, 2006.

[9] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[10] B. Julesz, *Foundations of Cyclopean Perception*. University of Chicago Press, 1971.

[11] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[12] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. D. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[13] J. Engel, J. Stückler, and D. Cremers, "Large-scale direct SLAM with stereo cameras," in *Proc 28th IEEE/RSJ Conf on Intelligent Robots and Systems*, pp. 1935–1942, 2015.

[14] S. Jung and C. J. Taylor, "Camera trajectory estimation using inertial sensor measurements and structure fom motion results," in *Proc 15th IEEE Conf on Computer Vision and Pattern Recognition*, vol. II, pp. 732–737, 2001.

[15] G. Nützi, S. Weiss, D. Scaramuzza, and R. Siegwart, "Fusion of IMU and vision for absolute scale estimation in monocular SLAM," *Journal of Intelligent and Robotic Systems*, vol. 61, no. 1-4, pp. 287–299, 2011.

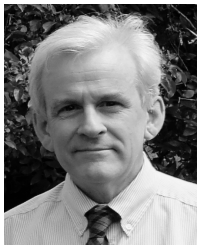
[16] G. Klein and D. W. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc 6th IEEE/ACM Int Symp on Mixed and Augmented Reality*, pp. 225–234, 2007.

[17] M. Achtelik, M. Achtelik, S. Weiss, and R. Siegwart, "Onboard IMU and monocular vision based control for MAVs in unknown in-and outdoor environments," in *Proc 2011 IEEE Int Conf on Robotics and Automation*, pp. 3056–3063, 2011.

[18] C. Hide, T. Botterill, and M. Andreotti, "Vision-aided IMU for handheld pedestrian navigation," in *Proc 23rd Int Technical Meeting of the Institute of Navigation (ION GNSS 2010) Portland, OR, Sep 21-24*, pp. 534–541, 2010.
[19] D. Scaramuzza, F. Fraundorfer, M. Pollefeys, and R. Siegwart, "Absolute scale in structure from motion from a single vehicle mounted camera by exploiting nonholonomic constraints," in *Proc 12th IEEE Int Conf on Computer Vision*, pp. 1413–1419, 2009.
[20] B. Kitt, A. Geiger, and H. Lategahn, "Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme," in *Proc Intelligent Vehicles Symposium III*, pp. 486–492, 2010.
[21] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *Proc IEEE Symposium on Intelligent Vehicles Symposium, IV2011*, pp. 963–968, 2011.
[22] S. Song, M. Chandraker, and C. C. Guest, "Parallel, real-time monocular visual odometry," in *Proc 2013 IEEE Int Conf on Robotics and Automation*, pp. 4698–4705, 2013.
[23] S. Song and M. Chandraker, "Robust scale estimation in real-time monocular SFM for autonomous driving," in *Proc 27th IEEE Conf on Computer Vision and Pattern Recognition*, pp. 1566–1573, 2014.
[24] T. Botterill, S. Mills, and R. D. Green, "Correcting scale drift by object recognition in single-camera SLAM," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1767–1780, 2013.
[25] H. Strasdat, J. M. M. Montiel, and A. J. Davison, "Drift aware large scale monocular SLAM," in *Proc Robotics: Science and Systems*, 2010.
[26] F. Lu and E. Milios, "Robot pose estimation in unknown environments by matching 2D range scans," *Journal of Intelligent and Robotic Systems*, vol. 18, no. 3, pp. 249–275, 1997.
[27] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g^2o: A general framework for graph optimization," in *Proc 2011 IEEE Int Conf on Robotics and Automation*, pp. 3607–3613, 2011.
[28] R. O. Castle, G. Klein, and D. W. Murray, "Combining monoSLAM with object recognition for scene augmentation using a wearable camera," *Image and Vision Computing*, vol. 28, no. 12, pp. 1548–1556, 2010.
[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[30] A. J. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *Proc 9th IEEE Int Conf on Computer Vision*, pp. II: 1403–1410, 2003.
[31] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 1052–1067, 2007.
[32] R. O. Castle and D. W. Murray, "Keyframe-based recognition and localization during video-rate parallel tracking and mapping," *Image and Vision Computing*, vol. 29, no. 8, pp. 524–532, 2011.
[33] J. Civera, D. Gálvez-López, L. Riazuelo, , J. D. Tardós, and J. M. M. Montiel, "Towards semantic SLAM using a monocular camera," in *Proc 24th IEEE/RSJ Conf on Intelligent Robots and Systems, San Francisco CA, USA, Sep 25-30, 2011*, pp. 1277–1284, 2011.
[34] S. Y. Bao and S. Savarese, "Semantic structure from motion," in *Proc 24th IEEE Conf on Computer Vision and Pattern Recognition*, pp. 2025–2032, 2011.
[35] N. Fioraio and L. Di Stefano, "Joint detection, tracking and mapping by semantic bundle adjustment," in *Proc 26th IEEE Conf on Computer Vision and Pattern Recognition*, pp. 1538–1545, 2013.
[36] D. Gálvez-López, M. Salas, and J. M. M. Montiel, "Real-time monocular object SLAM." arXiv:1504.02398 [cs.CV], 2015.
[37] A. Dame, V. A. Prisacariu, C. Y. Ren, and I. Reid, "Dense reconstruction using 3d object shape priors," in *Proc 26th IEEE Conf on Computer Vision and Pattern Recognition*, pp. 1288–1295, 2013.
[38] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.
[39] C. Bao, Y. Wu, H. Ling, and H. Ji, "Real time robust l1 tracker using accelerated proximal gradient approach," in *Proc 25th IEEE Conf on Computer Vision and Pattern Recognition*, pp. 1830–1837, 2012.
[40] H. Zhang, A. Geiger, and R. Urtasun, "Understanding high-level semantics by modeling traffic patterns," in *Proc 14th IEEE Int Conf on Computer Vision*, pp. 3056–3063, 2013.
[41] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, "3D traffic scene understanding from movable platforms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 1012–1025, 2014.
[42] P. J. Huber, *Robust Statistical Procedures: Second Edition*. Society of Industrial and Applied Mathematics, 1996.
[43] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc 9th European Conf on Computer Vision*, 2006.
[44] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
[45] E. Neufert and P. Neufert, *Neufert's Architects' Data*. Wiley-Blackwell, 2012.
[46] A. E. Cavelaars, A. E. Kunst, J. J. Geurts, R. Crialesi, L. Grötvedt, U. Helmert, , E. Lahelma, O. Lundberg, A. Mielck, N. K. Rasmussen, E. Regidor, T. Spuhler, and J. P. Mackenbach, "Persistent variations in average height between countries and between socio-economic groups: an overview of 10 European countries," *Annals of Human Biology*, vol. 27, no. 4, pp. 407–421, 2000.
[47] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.
[48] H. Strasdat, A. J. Davison, J. M. M. Montiel, and K. Konolige, "Double window optimisation for constant time visual SLAM," in *Proc 13th IEEE Int Conf on Computer Vision*, pp. 2352–2359, IEEE, 2011.
[49] M. Crocco, C. Rubino, and A. Del Bue, "Structure from motion with objects," in *Proc 29th IEEE Conf on Computer Vision and Pattern Recognition*, pp. 4141–4149, 2016.
[50] C. Rubino, M. Crocco, and A. Del Bue, "3D object localisation from multi-view image detections," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Duncan Frost gained a BSc and MSc in Electronic Engineering summa cum laude from the University of KwaZulu-Natal in 2010 and 2012 respectively, and a DPhil in Engineering Science from the University of Oxford in 2017. He is currently a postdoctoral researcher in Oxford's Active Vision Laboratory. His research interests are in real-time computer vision for augmented-reality applications.

Victor Prisacariu graduated with first class honors in computer engineering from Gheorghe Asachi Tehnical University in Romania in 2008, and gained the DPhil degree from the Department of Engineering Science, University of Oxford in 2012. He continued there first as an EPSRC prize post doctoral researcher, then as Dyson Senior Research Fellow before being appointed an Associate Professor in 2017. His research interests include semantic visual tracking, 3D reconstruction, and SLAM. He is a Research Fellow at St Catherine's College, Oxford.



David Murray graduated with first class honours in physics in 1977 and gained the DPhil degree in low-energy nuclear physics in 1980, both from the University of Oxford. He held a Research Fellowship in physics at the California Institute of Technology before moving to the General Electric Company's research laboratories in London, where he pioneered work on structure from motion, motion segmentation, and object recognition. He was appointed University Lecturer at Oxford in 1989 where he founded the Active Vision Laboratory. He was made a Professor of Engineering Science in 1997. His interests continue to lie in all aspects of motion understanding. He is a Fellow of St Anne's College Oxford, a Fellow of the IEE, and a member of the IEEE.