Planes, Trains and Automobiles – Autonomy for the Modern Robot

Gabe Sibley, Christopher Mei, Ian Reid and Paul Newman

***Abstract*—We are concerned with enabling truly large scale autonomous navigation in typical human environments. To this end we describe the acquisition and modeling of large urban spaces from data that reflects human sensory input. Over 200GB of image and inertial data are captured using head-mounted stereo cameras. This data is processed into a relative map covering 121 km of Southern England. We point out the numerous challenges we encounter, and highlight in particular the problem of undetected ego-motion, which occurs when the robot finds itself on-or-within a moving frame of reference. In contrast to global-frame representations, we find that the continuous relative representation naturally accommodates moving-reference-frames – without having to identify them first, and without inconsistency. Within a moving-reference-frame, and without drift-less global exteroceptive sensing, motion with respect to the global-frame is effectively unobservable. This underlying truth drives us towards relative topometric solutions like relative bundle adjustment (RBA), which has no problem representing distance and metric Euclidean structure, yet does not suffer inconsistency introduced by the attempt to solve in the global-frame.**

I. Introduction

Autonomous navigation in human working environments is an important problem, and this paper is motivated by our attempt to make sense of the 121 km path between Oxford and London depicted in Figs. 1 and 2. The map begins in an office in Oxford, and proceeds with various forms of transport including: foot, bicycle, train, subway, escalator, rickshaw, punting-boat and ferris wheel. Note that we cannot detect our true position in the global inertial frame – when we are traveling on the train or subway for instance, motion with respect to the global inertial frame becomes effectively unobservable in the presence of inertial sensing noise and drift. During this ~7 hour experiment we are able to compute relative metric motion estimates 89.4% of the time, falling back on a constant velocity model and inertial orientation sensing for the remainder.

This paper argues that a relative, topometric approach to autonomous navigation is not only sufficient, in the sense that one can find shortest paths in a map, but ultimately that it is *necessary* as well – that is, in order to solve many real world navigation tasks, we will have to adopt relative topological representations. This is in stark contrast to current received wisdom that it is possible – indeed preferable – to estimate everything in a single global coordinate frame.

This work has partly been supported by the European Commission under grant agreement number FP7-231888-EUROPA. This work has also been supported by the Systems Engineering for Autonomous Systems (SEAS) Defence Technology Centre.

—Gabe Sibley, Christopher Mei, Ian Reid and Paul Newman are with the University of Oxford, Department of Engineering Science, Parks Road, Oxford UK OX13PJ. {gsibley,cmei,ian,pnewman}@robots.ox.ac.uk

Figure 1: 121 km path between Oxford in the upper left and London in the bottom right. We compute visual estimates for 89.4% of this distance. Using appearance-based place recognition and inertial dead reckoning, 100% is covered topologically, which is sufficient for path planning. The graph begins in an office in Oxford, and proceeds with various forms of transport including: foot, bicycle, train, subway, escalator, rickshaw, punting-boat and ferris wheel. Note that we cannot detect our true position in the global inertial frame – when we are traveling on the train or subway for instance, motion with respect to the global inertial frame becomes effectively unobservable in the presence of noise.

First we describe our optimization framework that allows us to recover locally optimal structure using a relative metric manifold that describes the world. Note that it is not clear that it is *necessary* to estimate everything in a single coordinate frame – for instance most problems of autonomous navigation, such as path planning, obstacle avoidance or object manipulation, can be addressed within the confines of a *metric manifold*. Taking this route, we structure the problem as a graph of relative poses with landmarks specified in relation to these poses. In 3D this graph defines a connected manifold with a distance metric based on shortest paths. Notice that this is not a sub-mapping approach, as there are no distinct overlapping estimates, and there is only one objective function with a *minimal* parameter vector; similarly, this is not a pose-graph relaxation approach, as it solves for landmark structure as well. Our solution computes the optimal local Euclidean *metric* map structure throughout the relative manifold described by the continuous relative representation.

Second, we describe a large experiment in which 200GB of stereo data from a day trip to London are processed. No special care was taken to collect "clean" data that would lend itself to easy processing; the user collects data that reflect a typical human experience of the world (see Fig. 3). To highlight

Figure 2: Route around London with topologically interesting places for navigation. Between Paddington and Piccadilly the user is underground in the subway. From Piccadilly to Trafalgar Square and the London Eye the user is on foot. One loop was closed around the Eye. From the London Eye the user took the southern route West across the Thames, at which point he took a Rickshaw to Trafalgar Square and Piccadilly Circus. From Piccadilly Circus the user walked across Hyde Park to the Natural History Museum, at which point the batteries died, approximately 7 hours into the experiment.

this we compare statistics to data collected by our Segway robot, shown in Fig. 3. Processing such data from "the wild" is extremely challenging, and there are numerous difficulties encountered, which include but are not limited to: undetected ego-motion, motion blur, dynamic lighting changes, dropped frames, lens flare, dynamic obstacles, obstructed views, non-overlapping frames and power failures.

Third, we discuss the particularities of this data, and why it leads us to believe that topological methods are not only sufficient for navigation, they are the only way forward – that is they are also *necessary*. In particular, the problem of undetectable ego-motion deserves special attention as it ultimately forces the need for topological representations, and breaks all attempts to estimate a single global solution. Our data show that in real world situations the sensing vehicle is frequently *on* or *within* a moving frame of reference – with no recourse to global exteroceptive sensing. Traveling in a lift or on the subway are two examples. From our data alone, it is not possible, given the current state of the art, to infer a map in a single reference frame. Ultimately, we argue, it is also not desirable.

Fourth, to demonstrate a baseline *sufficiency* for navigation, we show that it is possible to find shortest paths in the relative maps we build – both in terms of time and distance. To demonstrate this, query images from Google of popular landmarks around London, such as the London Eye or Trafalgar Square, are matched to the the relative map to provide goals for route planning. This is particularly interesting as it allows users to give goals to robots in a natural way – by presenting them with images.

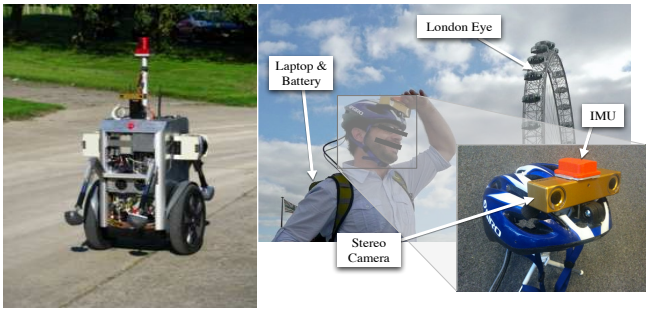


Figure 3: Processing data collected from human-like movement in urban spaces (right) is very different, and substantially more challenging than processing stable robot data (left). In both settings we capture 512x384 grey scale images at 20Hz.

We find that there are two fundamental components that enable navigation over large scales: the first and most important is reliable place recognition – we have to be able to close loops in a scalable fashion independent of any metric estimation; the second component is the wholesale adoption of a completely relative, topometric estimation framework. Beyond rudimentary route planning, processing this data so that it is useful for real-world navigation is a challenging task. Ultimately, we may need to learn to recognize commonly encountered moving-reference-frames, such as planes, trains and automobiles. Clearly, recognizing such high level semantic information is beyond the state of the art, and remains a challenge.

II. Related Work

Topological navigation is a well studied problem that was first addressed in robotics by Kuipers and Byun [23]. This work, and the later work by [4] seeks to describe interesting places as nodes in a graph of relative representations that encode metric information. In this context path planning is a matter of graph-search combined with local obstacle avoidance; based on this, topological representations have been used extensively for path planning. Recently numerous authors have recognized the benefits of vision for topological navigation and mapping [13][15][31][40], though often with the explicit goal of producing a globally embedded solution. To relate visual places, appearance-based recognition based on bag-of-words image matching is generally recognized as state-of-the art [6].

When it comes to metric vision-based estimation, bundle adjustment is the recognized optimal solution [12][22][42], though the problem is an old one [3][28]. The *full* problem tries to optimize the joint vehicle trajectory and map structure simultaneously given all measurements ever made. There are approximate incremental solutions that only optimize a small local subset of the map [8], and there are methods that approximate the full solution with various forms of marginalization [22][36], or by ignoring small dependency information [25][41]. Many use key-frames to reduce complexity, though at the expense of accuracy [10][21][29]. All these techniques suffer from computational complexity issues during loop closure.

In the context of long-term autonomy, roboticists recognize the need for online, real-time, navigation and mapping algorithms. This means that localization and mapping

algorithms must operate incrementally within a constant-time budget. Driven by this need, many authors have recognized the benefit of relative representations and manifolds [1][5][9][14][17][20][19][22][24]. On the other hand, the drawbacks of single-frame solutions have been recognized for some time [2].. The most common solution is probably sub-mapping [1][7][9][33], which breaks the estimation into many smaller mapping regions, computes individual solutions for each region, and then estimates the relationships between these sub-maps. Many difficult issues arise in sub-mapping, including map overlap, data duplication, map fusion and breaking, map alignment, optimal sub-map size, and consistent global estimation in a single Euclidean frame. In contrast, relative bundle adjustment is a *continuous* sub-mapping approach that avoids these complications.

The most successful global methods currently are the pose-graph optimization algorithms. Instead of solving the *full* problem, these methods optimize a set of relative pose constraints [16][32]. Like other methods, pose-graph solvers have worst-case complexity at loop closure that is dependent on the length of the loop.

The work most similar to our relative formulation is by Eade [9] and Konolige [22]. The former is akin to sub-mapping methods with constraints to enforce global Euclidean consistency at loop closure; the latter formulates the cost function relative to a single Euclidean frame and then makes a series of approximations to produce a sparse relative pose-graph. Neither method derives the purely relative objective function (incrementally, both rely on some form of privileged-reference frame), neither formulates the objective function completely without privileged frames, and both methods carry the burden of finding a globally consistent estimate in a single Euclidean frame. Our approach is substantially different because of the completely relative underlying objective function that we optimize.

III. Methods

In this section we first describe the continuous relative representation and how to optimize within this framework. Second we describe our robust stereo front-end processing pipeline. With this system we are able to achieve the kinds of *metric* accuracy shown in Fig. 4 and produce reconstructions like the one shown in Fig. 4.

A. Problem Formulation

This section re-caps the continuous relative representation (CRR) and relative bundle adjustment (RBA), which can be used to compute the MLE robot trajectory and map solution in a *relative* space [27][37]. Recall that bundle adjustment is the optimal global privileged-frame solution, in that its form matches the definition of the Cramer Rao Lower Bound (CRLB) [35]. However, BA quickly becomes too expensive (it is cubic in complexity), so we use a relative approach, which can be viewed as a *continuous* sub-mapping approach that runs with constant time complexity. BA seeks to minimize error between the observed and predicted measurements of n landmarks sensed from m sensor poses (or frames). Similarly, in RBA we minimize the difference between predicted and measured values. Let $l_{j,k}$, $k \in 1,...,n$, $j \in 1,...,m$ be a set of n 3D landmarks each parameterized relative to some *base-frame* j. Let t_j, $j \in 1,....,m$ be a set of m 6D relative pose relationships associated with edges in an undirected graph of frames. This graph defines a connected Riemannian manifold that is by definition everywhere locally Euclidean, though globally it is not embedded in a single Euclidean space. The relationship between parent-frame α and child-frame j is defined by a 4×4 homogeneous transform matrix, $T_{\alpha,j}$=$\hat{T}_{\alpha,j}T_{(t_j)}$, where $\hat{T}_{\alpha,j}$ is the current estimate and $T_{(t_j)}$ is the 4×4 homogeneous matrix defined by t_j. Each t_j parameterizes an infinitesimal delta transform applied to the relationship from its parent frame in the graph (i.e. an error-state formulation). The kinematic chain from frame j to frame i is defined by a sequence of 4×4 homogeneous transforms $T_{ji} = \hat{T}_{j,j+1}T_{(t_{j+1})}\hat{T}_{j+1,j+2}T_{(t_{j+2})},...,\hat{T}_{i-1,i}T_{(t_i)}$; the sensor model for a single measurement is thus

$$\begin{aligned} h_{i,k}(l_{j,k}, t_i, ...t_j) &= \mathcal{K}\left(T_{j,i}^{-1} l_{j,k}\right) \\ &= \mathcal{K}\left(g_{i,k}(l_{j,k}, t_{j+1}, ...t_i)\right) \end{aligned}$$

where $g_{i,k} : \mathbb{R}^{\dim(x)} \rightarrow \mathbb{R}^4$, $x \mapsto T_{j,i}^{-1} l_{j,k}$ transforms the homogeneous point $l_{j,k}$ from base-frame j to the observation frame i, and $\mathcal{K} : \mathbb{R}^4 \rightarrow \mathbb{R}^2$, is the standard perspective projection function [18]. This describes how landmark k, stored relative to base-frame j, is transformed into sensor frame i and then projected into the sensor. We make the usual assumption that measurements $z_{i,k}$ are independent and normally distributed: $z_{i,k} \sim N(h_{i,k}, R_{i,k})$. The cost function we associate with this formulation is

$$\begin{aligned} J &= \sum_{k=1}^{n} \sum_{i \in 1}^{m_k} \left(z_{i,k} - h_{i,k}(x)\right)^T R_{i,k}^{-1} \left(z_{i,k} - h_{i,k}(x)\right) \quad (1) \\ &\quad (m_k : \text{set of frames that see landmark } k) \quad (2) \\ &= \|z - h(x)\|_{R^{-1}}, \quad (3) \end{aligned}$$

which depends on the landmark estimate, $l_{j,k}$ and *all the transform estimates* $t_{j+1}, ...t_i$ *on the kinematic chain from the base-frame* j *to the measurement-frame* i. This problem is solved using iterative non-linear least-squares Gauss-Newton minimization for the values of x that minimize re-projection error — this yields the *maximum likelihood* estimate (subject to local minima). This process is slightly more expensive than traditional bundle adjustment, though the ultimate computational complexity is the same. An implementation detail is that to ensure scalability we have had to extend relative bundle adjustment with out-of-core graph processing, while still maintaining frame rate performance on the front end. At this point, the only limitation on the size of the maps we can build is storage capacity.

We stress that the relative solution is not equivalent to the normal Euclidean-space solution and it does not produce an estimate that can be easily embedded in a single Euclidean frame. Converting from the relative manifold into a single Euclidean space is a difficult problem that we argue is best handled by external resources that do not have constant run-

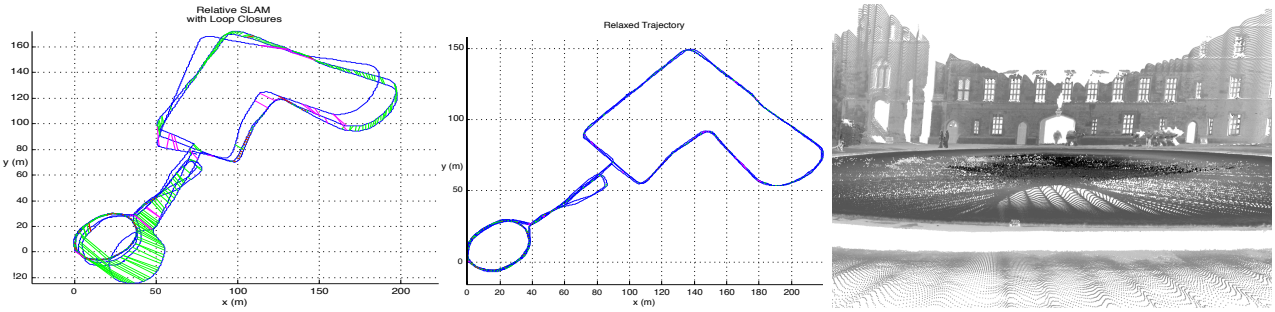


Figure 4: Before relaxation, with loop closures (left). After global pose graph relaxation (middle). Example of the metric pose estimate output by our system from the November 3, 2008 New College Data set [39]. This indicates that for small distances on the order of ~2km metric accuracy can be achieved to within 1-2m, and to within ~2cm after loop closure[37][27][30]. We posit this is sufficient for most path planning purposes. For example, see the Laser data rendered from the relative trajectory around New College Quad (right). Clearly, metric structure is available in the relative approach. In our experience this is sufficient for obstacle avoidance and local scene analysis.

time requirements - e.g. by operator computers, not on the robot.

B. Processing Pipeline

This section describes the engineering effort required to achieve precision, robustness and speed in the visual processing pipeline.

1) Image processing: includes rectification to allow fast scan line searches of corresponding left-right matches. Images are shifted to obtain the same mean and variance to aid left-right matching. We use FAST features [34] extracted at different levels of a scale-space pyramid for robustness to image blur. Detection thresholds are modified at each timestep to keep the number of detected features at a desired level independent of the scene.
2) Image alignment: An estimate of the 3-D rotation is obtained using the sum-of-squared-distance of image intensity using an efficient second-order gradient-descent minimization (ESM) as described in [26]. This greatly helps in cases with perceptual aliasing, such as bricks, tiles and picket fences.
3) Matching in time: The 3-D coordinates of the landmarks are projected into the left and right images and 9x9 patches are matched using mean shifted sum-absolute-difference error metric. Finally, ESM sub-pixel refinement is performed. Once matched, the current motion is estimated with a standard combination of RANSAC and a final robust m-estimation step [11][18].
4) Starting new landmarks: we typically track 100-150 features and use a multi-level quad-tree. At each pyramid level, a quad-tree captures how many features project into each cell. From these counts we can ensure an even spatial distribution across the image. Finally, upon initialization a SIFT descriptor is computed which can be used during re-localization and loop closure.

These steps help to ensure robustness to the challenging operating conditions illustrated in Figs. 6 and 7. Table I shows typical system performance results. Further details can be found in [27].

Feature Tracking			
	Avg.	Min.	Max.
Features per Frame	91	45	142
Feature Track Length	14.43	2	622
Re-projection Error	0.12	0.028	0.91

Table I: Typical tracking performance.

C. Loop Closure and Place Recognition

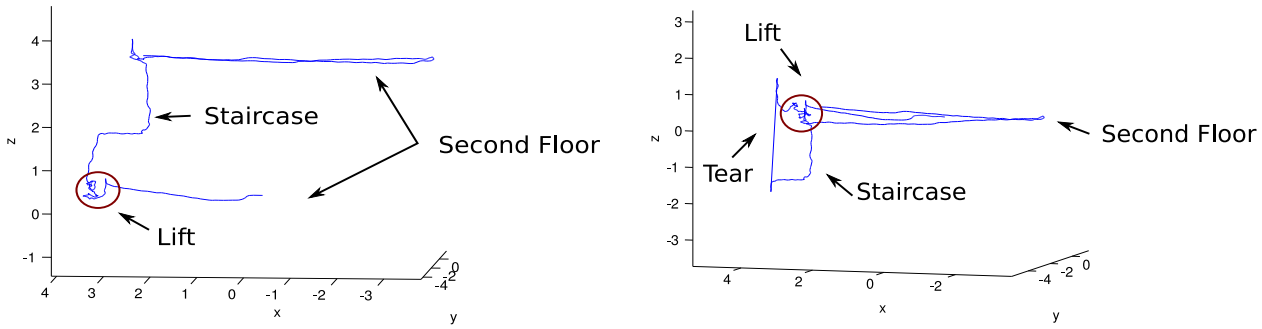
For loop closure and place recognition we rely on fast appearance-based mapping [6] which represents each place using the bag-of-words model developed for image retrieval systems in the computer vision community [38]. At time k the appearance map consists of a set of n_k discrete locations, each location being described by a distribution over which appearance words are likely to be observed. Incoming sensory data is converted into a bag-of-words representation; for each location, a query is made that returns how likely it is that the observation came from that location's distribution or from a new place. This allows us to determine if we are revisiting previously visited locations. In a filtering framework, incorrect loop closures are often catastrophic as the statistical estimates are corrupted. The CRR enables recovery from erroneous loop closures as removing the incorrect graph link and bad measurements returns the system to its previous state.

D. Path Planning

Presently path planning consists of finding shortest paths in the relative map with edges weighted either by distance or time. We use the magnitude of inter-frame motion (excluding orientation) to compute edge weights for distance based searches. Time-based searches use an edge weight that is simply the time between key frames, with an average value used for loop closure edges.

IV. Results

Our data consist of a largely unbroken stream of stereo measurements captured at 20hz, with relocalization across temporal gaps. The run to London consists of ~200GB of data spread over 6.9 hours. Every 50ms we update the relative motion estimate from visual methods, and fall back on the



(a) After exploring a floor, taking a flight of stairs followed by a lift, the robot returns to the same floor. No loop closure has currently been detected.

(b) A loop closure is triggered. The trajectory cannot be represented in Euclidean space and a "rip" appears (in this example in the staircase). Note that these rips are an artifact of embedding the solution in a single Euclidean space — they do not exist in the relative manifold. The size of the rip is related to the distance traveled in the lift. The topometric map is still usable for exploration because of its relative nature.

Figure 5: Lift sequence demonstrates the complexity of undetectable ego-motion with respect to a global-frame. Even with inertial sensing, moving-reference-frames like subways, lifts, trains etc. make the global-frame effectively unobservable given sensor noise.

Human Collected Data			
	Avg.	Min.	Max.
Distance Traveled (km)	—	—	121.4
Frames Processed	—	—	479726
Velocity (m/s)	1.3	0	8.2
Angular Velocity (deg/s)	8.6	0	191.5
Frames Per Second	33.7	21.3	46.4

Table II: Results for London human collected data. Note the difference in linear and angular velocity – this reflects the fact that head swivels result in very fast visual motion estimates. This type of motion is exactly the kind of challenge *not* faced in typical robot data.

Robot Collected Data			
	Avg.	Min.	Max.
Distance Traveled (km)	—	—	0.8
Frames Processed	—	—	29489
Velocity (m/s)	0.6	0	1.3
Angular Velocity (deg/s)	4.8	0	59.8
Frames Per Second	20.3	7.4	28.6

Table III: Robot data collected on a Segway RMP (see Fig. 3). Note the difference in linear and especially angular velocity in Table II.

IMU or a constant velocity motion model in the absence or failure of visual tracking. This strategy succeeds with visual estimation 89% of the time, and falls back on dead-reckoning for the remainder, giving an unbroken linear chain of relative relationships.

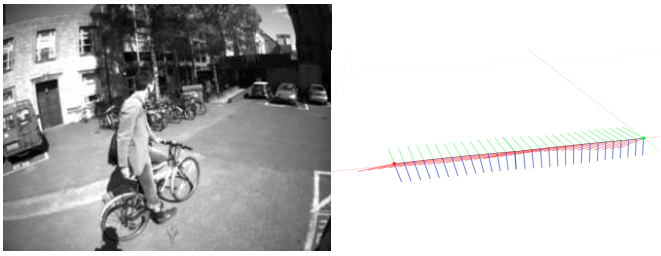
To illustrate the metric accuracy of the local estimation engine see Fig. 4, which shows an accuracy of ~1m after 2km of travel. Such performance is achieved via the combination of sub-pixel refinement, multi-level matching, spreading features across the image within a quad-tree, and exploiting loop closures to re-localize against the map [27][37].

We adopt the view that closures produced by FABMAP can be viewed as pinch-points between disparate points along this chain [30]. Note however, that a consistent global representation is impossible to compute in the presence of undetected ego-motion, such as the lift example in Fig. 5 or the numerous cases shown in Fig. 7. Given noisy inertial sensing, any attempt to compute a global solution is doomed to failure as the measurements are inconsistent with a single Euclidean embedding.

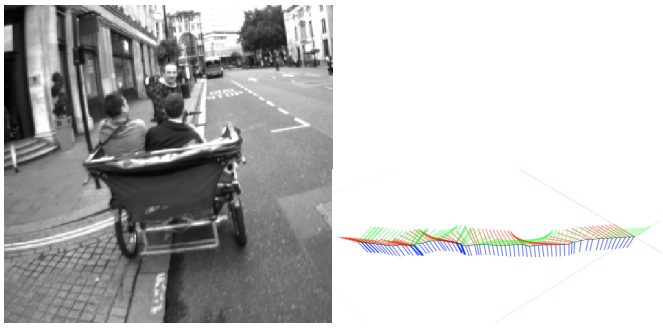
A. *Path Planning*

To plan a route in the graph we begin with a Google image from Trafalgar Square matched to an image from the graph with FABMAP (for example see Fig. 8). Given this query, we can find a path from Oxford to London based on either metric, temporal or topological distance. Note however that information can also go in the other direction – that is, given such appearance-based matches from the Internet, it is possible to label relative map with search words describing places.

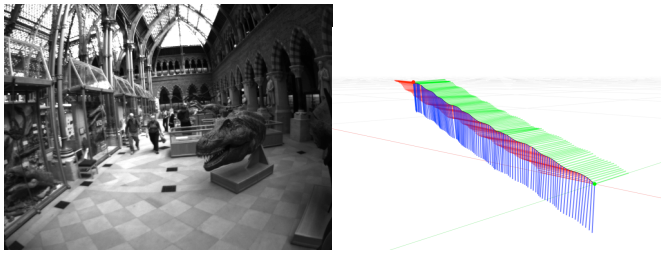
In the relative representation, shortest paths in the graph can be computed in a variety of ways; for instance, we might ask for the shortest path metrically, with respect to time, or purely topologically. Two paths from the Natural History Museum in Oxford to the London Eye are computed. The first path, based on a desired shortest time travel, takes the *southern* bridge from West to East (see Fig. 2). This route is shorter due to the rickshaw the user rode during exploration. The second path which gives shortest distance, traverses the loop closure at Trafalgar Square, and then takes the *northern* bridge to the London eye, which was traversed on foot and is indeed the more direct route (see Fig. 9 for representative loop closures). Clearly, paths planned in the relative representation can take advantage of metric and temporal information.



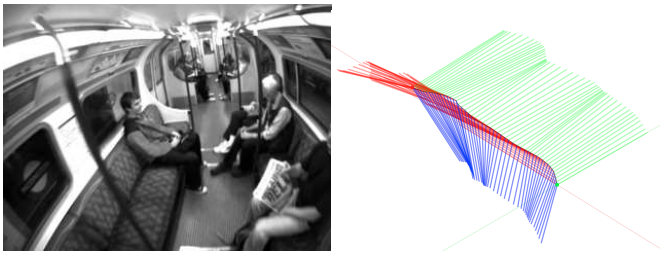
(a) A smooth constant velocity trajectory from a bicycle.



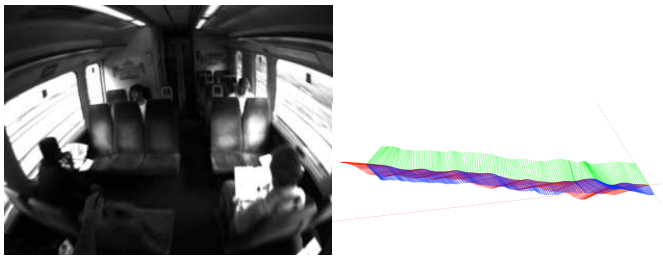
(b) Track estimated from the rickshaw showing user head swivel. Not as fast as the bicycle.



(c) Walking in the Oxford Natural History Museum. Note the clearly visible gait.

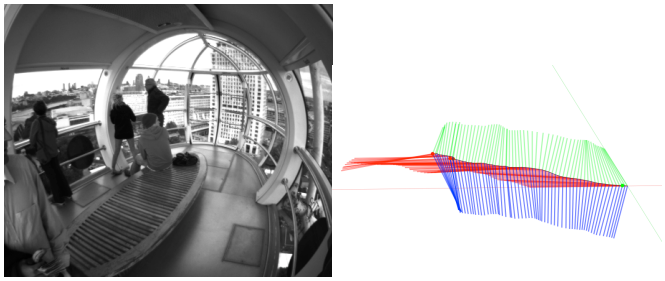


(d) Walking into a subway car. Not that at some point in this trajectory, the car beings to move, a fact not visually discernible here. Detecting linear acceleration with off-the-shelf inertial sensors is difficult in this situation.

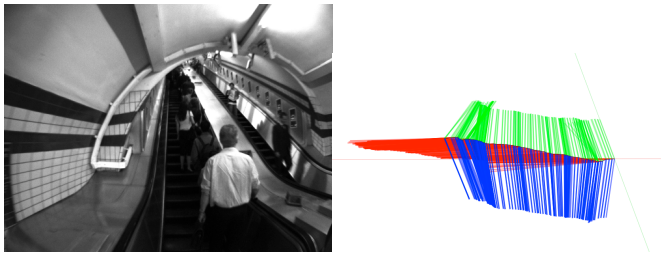


(e) Walking on a train looks like walking anywhere else. Note the extreme motion blur in the windows from the quickly passing external terrain.

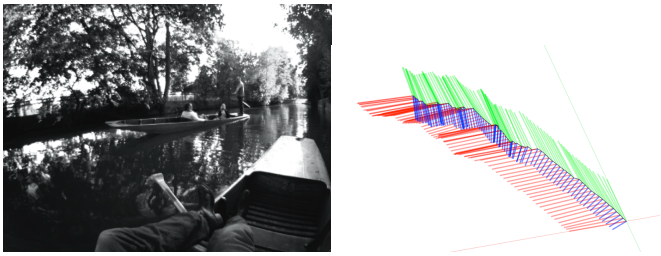
Figure 6: Excerpts from the London experiment show typical visually estimated trajectories for various modes of transport.



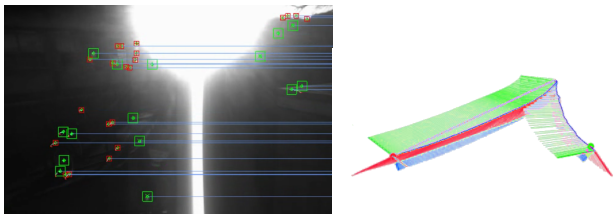
(a) Walking on the London Eye. It is difficult to detect loop closure metrically in the trip around the wheel.



(b) Extremely challenging escalator ride. Tracked motion oscillates between moving and stationary (note the bunched axes where it is stationary). Detecting linear acceleration with off-the-shelf inertial sensors is difficult in this situation.



(c) Visual motion estimation while punting is challenging due to reflections which appear to cause a slight instability.



(d) Successful feature tracking under challenging lighting conditions.

Figure 7: More trajectories for various modes of transport.

B. *Undetected Ego-Motion*

In practice, undetectable ego-motion due to inertial drift is a common phenomenon – as illustrated in the lift example in Fig. 5 or the numerous cases shown in Fig. 7. Navigation in the real world frequently travels on and inside various forms of moving-reference-frames, without the ability to accurately sense the global-frame. In the challenging motions shown in Figs. 6 and 7 note that walking is always similar, even though it is often taking place within a moving-reference-frame, such as the Tube, a train, or the London Eye. Fig. 7 in particular shows trajectories estimated from sequences on an escalator, in a boat, and under harsh lighting conditions. Tables II and III compare the London dataset to data collected on a Segway

Figure 8: Google image from Trafalgar Square (left) and matched image from the graph (right). Given this query, we find a path from Oxford to London based on either metric, temporal or topological distance. Note however that information also goes in the other direction – that is, given such appearance-based matches from the Internet, it is possible to label relative map with search words describing places. This is not surprising, especially for highly distinctive places like the London Eye, Piccadilly Circus, Trafalgar Square or the Natural History Museum [38].

Figure 9: Loop closure candidates at Trafalgar Square and Piccadilly Circus. Loops closure allows finding potentially shorter paths in the map, though of course such co-observability does not guarantee traversability.

RMP, which highlights the difficulty in processing human-like sensory input.

V. Discussion

Algorithms that solve for robot position in the privileged inertial coordinate frame are very different from relative approaches – they have different objective functions and they solve for different quantities. Privileged-frame solutions seek to embed the entire robot trajectory in a single Euclidean space; relative solution solves in a connected Riemannian manifold. The relative manifold is a metric space, and distance between two points can be computed from shortest paths in the graph. We have shown that the relative representation is amenable to planning (because path planning algorithms are commonly defined over graphs). Further, because the manifold is (by definition) locally Euclidean, we have access to highly accurate local metric structure at any time (for instance see Fig. 4 which shows lasers rendered from the relative trajectory around New College Quad). We posit that topometric solutions are not only sufficient for real world navigation, they are increasingly necessary. For instance, given undetectable ego-motion, it is not possible to build consistent map structures in a privileged-frame on which to navigate. We have endeavored to show that it is possible within a purely relative approach.

While it is certainly possible to build large scale, consistent global world models (especially with the use of GPS), we find that there are numerous real world situations where it is impossible to do so – that is, even GPS is not sufficient. There is no such thing as drift-free inertial sensing, and there are many examples in which position in the global inertial frame is effectively unobservable – places like lifts, subways, trains – and these are places where we want to navigate autonomously. This fact bears scrutiny, and helps us focus on much harder problems we will have to solve in order to move forward. These are problems such as learning when and where undetectable ego-motion becomes probable – that is, automatically discovering the location of transportation portals. It is interesting that one solution appears to be learning to recognize high-level semantic objects, such as lifts, escalators, planes, trains and automobiles. Given such labels then perhaps we can relate the topometric world to the global inertial frame. But the question remains – why? Certainly, if it is possible to embed useful metric information into a relative approach, and if it suffices completely for autonomous navigation, then why should we seek a global embedding?

While we claim that a topological approach is necessary for large scale real world navigation, note that our claims of navigational-path-planning-sufficiency are based on the assumption that co-observability implies traversability, and that we know how to handle the various modes of transport encountered during the traversal of paths. So, while it may be sufficient from the pure graph-search point of view, it is clearly limited because of the lack of higher-level knowledge about moving-reference-frames – that is, path planning along routes that include train, lifts, etc, must be informed about the temporal schedule of the trains, lifts, etc – it needs to know how to board a train or a lift. This is a substantially harder problem, and one that we argue is necessary to solve if we are going to effectively use these transportation modes for autonomous navigation. Presently, we side step this problem by simply defining sufficiency in terms of the ability to find shortest paths in a graph – i.e. we use the traditional definition, even though it is no longer appropriate once we have moving-reference-frames. This is interesting grounds for future research.

VI. Conclusion

In the presence of moving-reference-frames, and in the absence of drift-free global exteroceptive sensing, motion with respect to the inertial-frame can go undetected. We posit that this fact severely limits the scalability and applicability of single-coordinate-frame SLAM approaches. Undetectable ego-motion is a common phenomenon encountered in human operating environments — for instance on trains, planes, automobiles, lifts, etc. This underlying fact motivates relative topometric solutions like relative bundle adjustment, which has no problem representing distance and metric Euclidean structure, yet does not suffer inconsistency introduced by the attempt to solve in a global-frame. In a continuous relative formulation we find that moving-reference-frames can be accommodated without having to identify them first, and

without inconsistency. The relative representation is tolerant to structures that are frequently encountered in real situations. To explore the feasibility and scalability of our approach, over 200GB of image and inertial data are processed to produce relative estimates covering 121 km of Southern England. We describe the acquisition and modeling of large urban spaces from data that reflect a typical human experience of the world. We point out the numerous challenges that ensue, and highlight in particular the problem of undetected ego-motion, which is encountered when the robot finds itself on or within a moving frame of reference. We are concerned with enabling truly large scale autonomous navigation in typical human environments. In stark contrast to current received wisdom that it is possible – indeed preferable – to estimate everything in a single global coordinate frame, we argue that a relative, topometric approach to navigation is not only sufficient, but that it is necessary as well.

REFERENCES

[1] M. C. Bosse, P. M. Newman, J. J. Leonard, and S. Teller. SLAM in large-scale cyclic environments using the atlas framework. *International Journal of Robotics Research*, 23(12):1113–1139, December 2004.

[2] R. Brooks. Visual map making for a mobile robot. In *IEEE International Conference on Robotics and Automation*, 1985.

[3] D.C. Brown. A solution to the general problem of multiple station analytical stereotriangulation. Technical report, RCP-MTP Data Reduction Technical Report No. 43, Patrick Air Force Base, Florida (also designated as AFMTC 58-8), 1958.

[4] H. Choset and K. Nagatani. Topological simultaneous localization and mapping (SLAM): toward exact localization without explicit localization. *IEEE Transactions on Robotics and Automation*, 17:125–137, 2001.

[5] L. A. Clemente, A. J. Davison, I. Reid, J. Neira, and J. D. Tardos. Mapping large loops with a single hand-held camera. In *Robotics: Science and Systems*, 2007.

[6] M. Cummins and P. Newman. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *International Journal of Robotics Research*, 27(6):647–665, 2008.

[7] A. Davison, I. Reid, N. Molton, and O. Stasse. MonoSLAM: Realtime single camera SLAM. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 29(6):1113–1139, 2007.

[8] M. C. Deans. *Bearings-Only Localization and Mapping*. PhD thesis, School of Computer Science, Carnegie Mellon University, 2005.

[9] E. Eade and T. Drummond. Unified loop closing and recovery for real time monocular SLAM. In *Proceedings British Machine Vision Conference*, September 2008.

[10] C. Engels, H. Stewenius, and D. Nister. Bundle adjustment rules. In *Photogrammetric Computer Vision*, 2006.

[11] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381 – 395, 1981.

[12] A. W. Fitzgibbon and A. Zisserman. *Automatic Camera Recovery for Closed or Open Image Sequences*. Springer, 2004.

[13] F. Fraundorfer, C. Engels, and D. Nister. Topological mapping, localization and navigation using image collections. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007.

[14] U. Frese and T. Duckett. A multigrid approach for accelerating relaxation-based SLAM. In *Proceedings IJCAI Workshop on Reasoning with Uncertainty in Robotics (RUR 2003)*, pages 39–46, Acapulco, Mexico, 2003.

[15] T. Goedem'e, M. Nuttin, T. Tuytelaars, and L. V. Gool. Omnidirectional vision based topological navigation. *International Journal of Computer Vision*, 74(3):219?236, 2007.

[16] G. Grisetti, C. Stachniss, S. Grzonka, and W. Burgard. A tree parameterization for efficiently computing maximum likelihood maps using gradient descent. In *Proceedings Robotics: Science and Systems*, 2007.

[17] J.E. Guivant and E.M. Nebot. Optimization of the simultaneous localization and map-building algorithm for real-time implementation. *IEEE Transactions on Robotics and Automation*, 17(3):242–257, June 2001.

[18] R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[19] A. Howard. Real-time stereo visual odometry for autonomous ground vehicles. In *IEEE Conference on Robots and Systems (IROS)*, 2008.

[20] A. Howard, G. S. Sukhatme, and M. J. Mataric. Multirobot simultaneous localization and mapping using manifold representations. *Proceedings of the IEEE*, 94(7):1360–1369, 2006.

[21] G. Klein and D. Murray. Improving the agility of keyframe-based SLAM. In *European Conference on Computer Vision*, 2008.

[22] K. Konolige and M. Agrawal. FrameSLAM: from bundle adjustment to realtime visual mapping. *IEEE Transactions on Robotics and Automation, IEEE Journal of Robotics and Automation, International Journal of Robotics Research*, 24(5):1066–1077, 2008.

[23] B.J. Kuipers and Y.T. Byun. A robust qualitative method for spatial learning in unknown environments. In *Proc. National Conference. of Artificial Intelligence*, 1988.

[24] A. Martinelli, V. Nguyen, N. Tomatis, and R. Siegwart. A relative map approach to SLAM based on shift and rotation invariants. *Robotics and Autonomous Systems*, 55(1):50–61, 2007.

[25] P. F. McLauchlan. The variable state dimension filter applied to surface-based structure from motion. Technical report, University of Surrey, 1999.

[26] C. Mei, S. Benhimane, E. Malis, and P. Rives. Efficient homography-based tracking and 3-d reconstruction for single-viewpoint sensors. *IEEE Transactions on Robotics and Automation*, 24(6):1352–1364, 2008.

[27] C. Mei, G. Sibley, M. Cummins, I. Reid, and P. Newman. A constant-time efficient stereo SLAM system. In *British Machine Vision Conference*, 2009.

[28] E. M. Mikhail. *Observations and Least Squares*. Rowman & Littlefield, 1983.

[29] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyse, and P. Sayd. Real time localization and 3d reconstruction. In *Proceedings of Computer Vision and Pattern Recognition*, 2006.

[30] P. Newman, G. Sibley, M. Smith, M. Cummins, A. Harrison, C. Mei, I. Posner, R. Shade, D. Schroeter, L. Murphy, W. Churchill, D. Cole, and I. Reid. Navigating, recognizing and describing urban spaces with vision and lasers. *International Journal of Robotics Research*, 1:1–28, 2009.

[31] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–659, Washington, DC, 2004.

[32] E. Olson, J. Leonard, and S. Teller. Fast iterative alignment of pose graphs with poor initial estimates. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2262–2269, 2006.

[33] P. Pinies and J. D. Tardos. Scalable slam building conditionally independent local maps. In *IEEE conference on Intelligent Robots and Systems*, 2007.

[34] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *European Conference on Computer Vision*, May 2006.

[35] G. Sibley. *Long Range Stereo Data-fusion from Moving Platforms*. PhD thesis, University of Southern California, 2007.

[36] G. Sibley, L. Matthies, and G. Sukhatme. *A Sliding Window Filter for Incremental SLAM*, chapter 7, pages 103–112. Springer Lecture Notes in Electrical Engineering, 2007.

[37] G. Sibley, C. Mei, I. Ried, and P. Newman. Adaptive relative bundle adjustment. In *Robotics: Science and Systems*, 2009.

[38] J. Sivic and A. Zisserman. Video google: Efficient visual search of videos. In *Toward Category-Level Object Recognition*, pages 127–144, 2006.

[39] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman. The new college vision and laser data set. *International Journal of Robotics Research*, 28(5):595–599, 2009.

[40] B. Steder, G. Grisetti, S. Grzonka, C. Stachniss, A. Rottmann, and W. Burgard. Learning maps in 3d using attitude and noisy vision sensors. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007.

[41] S. Thrun, D. Koller, Z. Ghahmarani, and H. Durrant-Whyte. SLAM updates require constant time. In *Workshop on the Algorithmic Foundations of Robotics*, December 2002.

[42] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – A modern synthesis. In W. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, LNCS, pages 298–375. Springer Verlag, 2000.