Real-Time Bounded-Error Pose Estimation for Road Vehicles Using Vision

Ashley Napier, Gabe Sibley and Paul Newman
Mobile Robotics Group University of Oxford

***Abstract*—This paper is about online, constant-time pose estimation for road vehicles. We exploit both the state of the art in vision based SLAM and the wide availability of overhead imagery of road networks. We show that by formulating the pose estimation problem in a relative sense, we can estimate the vehicle pose in real-time and bound its absolute error by using overhead image priors. We demonstrate our technique on data gathered from a stereo pair on a vehicle traveling at 40 kph through urban streets. Crucially our method has no dependence on infrastructure, needs no workspace modification, is not dependent on GPS reception, requires only a single stereo pair and runs on an every day laptop.**

I. Motivation and Background

It is hard to understate the importance of the transport of goods and people in daily life. We are totally dependant on it, thus, any increase in efficiency, access, safety or reliability will have a major economic and societal impact. This paper describes work towards this goal, motivated by the belief that information engineering, computing and robotics can provide a low cost solution for smart vehicles in civil, defence and industrial domains. Such vehicles offer the possibility of end-to-end goods transportation, improved efficiency and safety on our roads, and give our aged, infirm and sensorially impaired citizens the hope of independent personal transport.

A foundation technology for smart vehicles is accurate online pose estimation. Currently, this is achieved using dedicated navigation infrastructure such as GPS in outdoor environments and markers in warehouses etc. Robots in factories generally rely on accurate placement of underground cables or reflecting beacons; this is expensive and inconvenient, but admissible because the workspace is small. However, the cost of marking *every* space in which we desire machines to navigate; cities, highways, public buildings, hospitals, warehouses, building perimeters, docks, airports, mines, etc - is utterly prohibitive[1]. Systems using GPS are subject to blocked or sporadic signals. Our aim is to remove the dependance on such infrastructure - our goal is *infrastructure-free vast scale navigation*. It is important to understand the importance of the phrase *'infrastructure-free'*. If by judicious use of data from vehicle-mounted sensors, a machine could navigate unaided on our roads then a whole new vista of opportunities opens up, bringing with it major commercial and social benefits.

We assume that a road vehicle can access, via web search or disk access when web access is unavailable, overhead images of its current location. Furthermore, we assume that modest onboard processing is available - a 2GHz laptop equivalent. We require nothing of the workspace - no prior survey, no beacons, no dedicated infrastructure. We require no modifications of the vehicle other than the mounting of a small OEM stereo pair. We do not require awkward interfacing with an inertial or odometry system. Our system is very much plug and play. Importantly we can operate when and where GPS signals are blocked or sporadic and in principal at precisions commensurate with GPS.

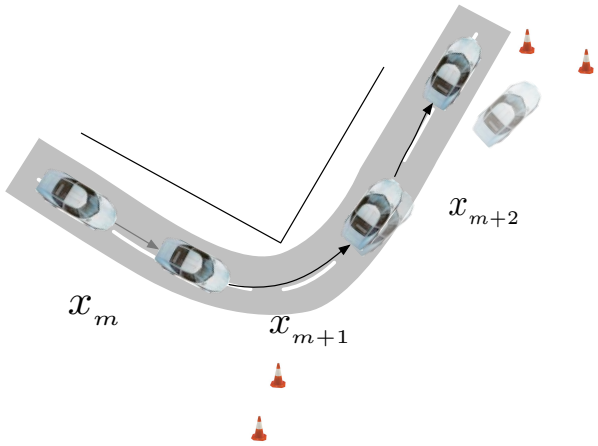

Figure 1: An illustration of the consequences of the RBA formulation. Here observations of features after turning necessitate a large perturbation in the relative transformation x_{m+2}. Note that x_m is largely unaffected particularly because none of the features seen before cornering are visible afterwards.

Our method leverages recent work on relative SLAM using vision, pose-graph optimisation and exploitation of prior knowledge in the form of aerial images. We are not the first to think along these lines. Work by Kummerle et. al [5] has used overhead imagery to correct SLAM-derived vehicle trajectories but this relied upon a strict correlation between edges in an aerial image and vertical surfaces perceived by laser scanners rather than vision (see also [11] for a survey of previous low precision matching techniques employed in transport research). Our method does not require vertical surfaces or a laser. Instead it uses the visual appearance of the scene to perform the corrections to a SLAM derived trajectory. Like us, Levinson et. al [6] considered using offline constructed reflectance imagery to provide constraints in a pose-graph adjustment but they do so in a single global frame which prohibits online operation. This is an important distinction,

[1]Nevertheless this has been the approach taken by several European transport projects for example [2], [1] .

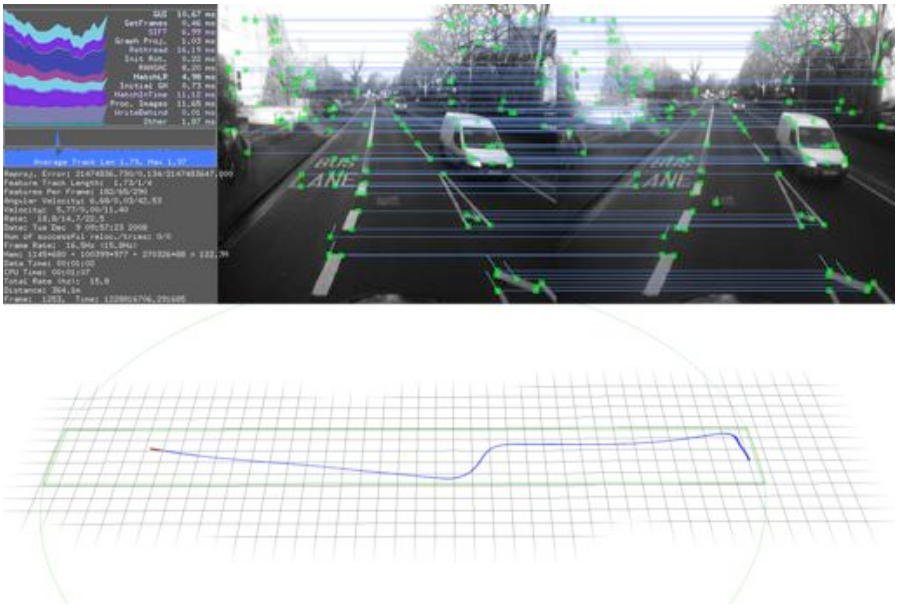

Figure 2: The typical RBA debugging/visualisation display provided to the user giving the reader a sense of the urban workspace we are operating in and its scale. Here the vehicle is ~1.1km into the 100km "Highway" data set. The Figure shows the current stereo image pair with correspondences between features shown in blue. Features are also tracked over time and this association history is indicated as commet tails on each feature. The lower half of the figure shows part of the vehicle trajectory (a few kilometres here) rendered for visualisation's sake into a single metric frame.

above all it is the particular relative formulation of our work which affords us the opportunity to pursue a large scale online implementations.

This paper is structured as follows Section II explains how using a stereo camera we can create a locally accurate metric picture of the vehicle motion. Section III describes how this very accurate estimate of recent vehicle motion can be used to synthesise an aerial view image of the vehicles immediate surroundings. Synthetic images are used in a registration process to yield metric constraints which, as described in Section IV, are used to produce a local correction to the estimated trajectory. Section V presents some early results of our technique before Section VI concludes with a discussion of future work.

II. Pose Graph Generation with RBA

This paper draws on our previous work on relative bundle adjustment (RBA) which has resulted in a frame-rate pose estimation system driven by stereo image pairs. A detailed summary of the system can be found in [9] but we are well served by a synopsis here. The fundamental idea is that massive computational savings can be achieved by computing the pose of a vehicle not in a single global and hence privileged frame, but on a manifold which locally has high metric precision but has only topological correctness at vast scales. RBA formulates the SLAM problem as the task of solving for vehicle and feature states given a sequence of observations (for example from a stereo camera) - this then is no different from conventional metric SLAM. Also, in common with formulations like GraphSLAM [10], we solve for the pose of a chain of vehicle poses - essentially one vehicle pose for every time step. The big difference between RBA and standard SLAM techniques stems from the fact that in RBA features and poses are not registered in a common frame. Instead features are represented relative to one of the individual vehicle poses and vehicle poses are defined relative to one another and *not* relative to a common origin. At run time stereo derived measurements between vehicle poses and features are used to minimise a re-projection error cost function. We stress again that it is the *relative* vehicle poses and features states that are being perturbed so that the total re-projection error of features into all cameras frames is minimised. It is important to understand how this formulation yeilds constant time operation. Consider the case in which, in the light of a particular measurement, a proportional correction must be made to a transformation between a pair of vehicle poses. This correction will also impinge on neighbouring poses and features leading to additional perturbations. Of course, these corrections in turn will fan out through the state vector to other poses and features but crucially they will do so with ever diminishing magnitude. To understand why this is so consider Figure 1. Here observations of features after turning a corner necessitate a perturbation in the relative transformation x_{m+2}. Note however that x_m is largely unaffected - particularly because none of the features seen before the corner are visible afterwards. Mathematically x_{m+2} and x_m are coupled - a perturbation to one influences the other. The efficiency comes from the observation and exploitation of the fact that the effects of coupling between x_m and x_{m+i} drop quickly - as $1/|i|$. Thus at run time one can pose the optimisation only over poses undertaking numerically meaningful adjustments. Typically this implies involvement of under twenty relative poses. The central point here is that although the totality of relative chain can grow without bound, at run time only a subset is involved in active optimisation.

Our current relative SLAM engine runs at framerate (20-

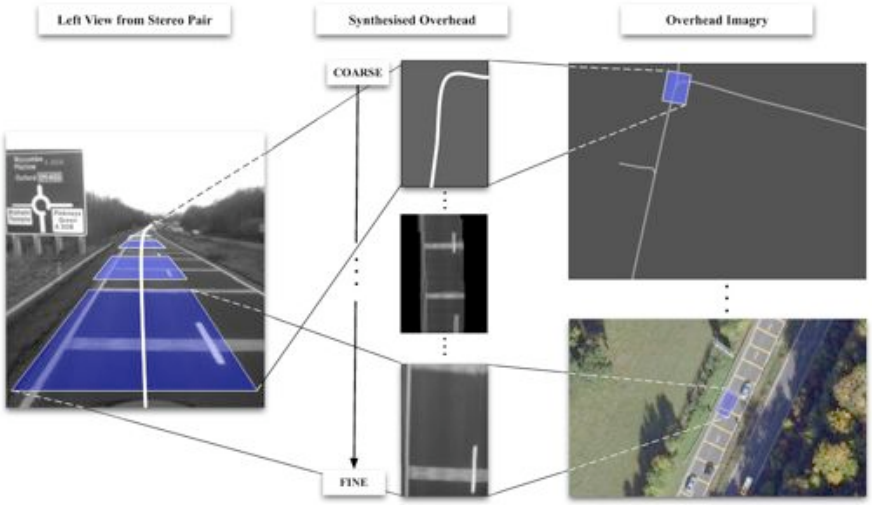

Figure 3: Overview of the coarse-to-fine road matching scheme. During nominal operation the vehicle is well localized in pixel coordinates in the overhead image. Using a coarse-to-fine search strategy, the minimum cost image-to-image alignment is found using robust efficient second-order minimization (ESM) [7].

40Hz) on 512 by 384 mono images delivered by a Point Gray Bumblebee camera. We use a parallel-tracking-and-mapping framework to ensure realtime operation [4]. Figure 2 shows the typical debugging display provided to the user. It is included here to give the reader a sense of the urban workspace we are operating in and its scale. The Figure shows the current stereo image pair with correspondences between features shown in blue. Features are also tracked over time and this association history is indicated as comet tails on each feature. The lower half of the figure shows part of the vehicle trajectory (a few kilometres here) rendered for visualisation's sake into a single metric frame.

III. Relating the Ground View To the overhead prior

In this Section we will describe how we utilise overhead images[2] in conjunction with RBA derived pose-graphs to yield constraints that will be used in Section IV to produce a constant-time estimate of the vehicle location relative to a road centered coordinate frame. The proceedure has two distinct steps - view synthesis followed by overhead/synthetic image registration.

A. Local Overhead View Synthesis

We assume at the outset that we know where the vehicle is initially in pixel coordinates inside an aerial image $\mathcal{I}_\pi$. The RBA system produces an endless chain of relative vehicle poses $X = \{x_0, x_1 \cdots\}$. We select $\overline{X} \subset X$, over the last D meters of trajectory such that $\overline{X} = \{x_m, x_{m+1} \ldots x_i\}$. Each pose $x_p \in \overline{X}$ has a pair of stereo images $[I_L, I_R]$ associated with it. Now, the RBA system itself has detected, inferred and made use of 3D point features visible in these images. For each image pair and corresponding 3D point features we use RANSAC to estimate a ground plane for the area immediately in front of the vehicle. This ground plane estimation allows us to generate a homography to warp the camera view to an orthonormal "birds-eye" view of the ground scene immediately infront of the vehicle, we only use the area directly infront of the vehicle to avoid non-road pixels. Furthermore, because we have a sequence of relative poses and images in $\overline{X}$ we can synthesise an extended super-resolution birds-eye image $\mathcal{I}^+$ over the last D meters of trajectory. Note that via ground plane detection and warping, we have implicitly ironed flat a section of the relative manifold created by RBA.

B. Ground View Matching

Given our synthesised image $\mathcal{I}^+$ we can now try to match it against the contents of $\mathcal{I}_\pi$. Our aim is to produce a $SE2$ transformation between the pose at the head of the pose chain in $\bar{X}$, and a frame on the road.

During nominal operation the vehicle is well localized in pixel coordinates in the overhead image. Using a coarse-to-fine search strategy (see Figure 3), the minimum cost image-to-image alignment is found using robust efficient second-order minimization (ESM) [7]. This ensures good matches whenever they are possible at the highest resolution (individual road tiles), and falls back to higher levels (road structure) when the high-res matching fails. Note that we perform the match in pixel space and the conversion factor (pixels/m) between image and metric space applied throughout is dictated by $\mathcal{I}_\pi$.

IV. Local Pose-Graph Relaxation

The RBA subsystem produces a chain of 6DOF (Degrees Of Freedom) vehicle poses linked by relative transformations which should be thought of as uncertain metric constraints. In this section we will describe how constraints obtained from the processing described in Section IIIcan be fused with this pose-graph in a secondary processing step. The result is a vehicle trajectory segment which is constrained to lie on the driven road which, unlike a standalone RBA implementation, admits queries about the vehicle pose relative to the road itself.

[2]Overhead images from www.getmapping.com

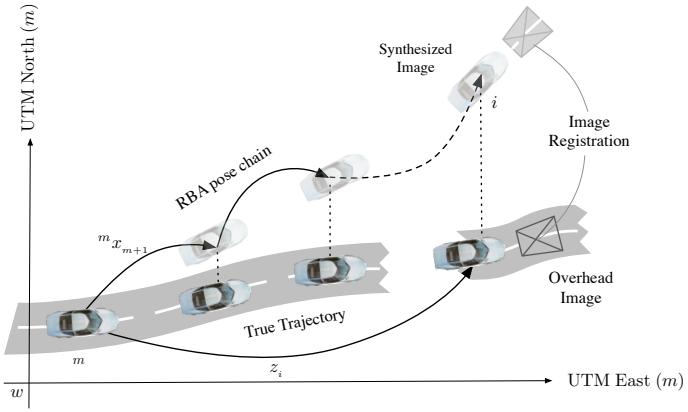

Figure 4: Problem setup showing accumulated error in RBA and ground view matches used for correction.

Note that we will end up with pose estimates possessing all the 6DOF precision of the RBA system (sub cm) but with any large-scale drift removed by our recourse to the overhead image implicit in the use of constraints produced in Section III.

We wish to "relax" the pose-graph, perturbing the edges to accommodate, in a minimum error sense, the metric information in both RBA-derived (inter-pose) and overhead constraints. Several authors have examined methods for pose-graph relaxation in recent years e.g. [10], [8], [3]. The particular size and simple structure of our graphs motivates us to use classical non-linear optimisation techniques taking care at implementation time to make full use of the sparse properties of the problem.

Consider a kinematic chain of relative poses, for example as shown in 4. Let $x_i \in SE_2$ be the relative-transform, or edge, from pose $i-1$ to pose i, and let $\hat{x}_i \sim \mathcal{N}(x_i, \Pi_i)$ be a random variable capturing prior knowledge about x_i. Here Π_i is a covariance matrix expressing uncertainty in the relative transformation x_i which can be provided by the RBA system. We will write ${}^{m}x_i \in SE_2$ when referring to a chain of relative transforms from pose m to pose i. Pose m is the anchor pose - it is an initial estimate given in the UTM (Universal Transverse Mercator) frame of the overhead image.

Now consider the case in which an implementation of the pipeline in Section III has produced a measurement z_i which relates frame i to frame m in the overhead prior image. If $h_i(x) = {}^{m}x_{m+1} \oplus {}^{m+1}x_{m+2} \oplus, ..., \oplus {}^{i-1}x_i$, is the integrated kinematiac chain from m to i, then we can write the measurement error as

$$\begin{aligned} z_i &= {}^{m}x_{m+1} \oplus {}^{m+1}x_{m+2} \oplus, ..., \oplus {}^{i-1}x_i + v_i \\ &= h_i(x) + v_i \end{aligned} \quad (1)$$

where $v_i \sim \mathcal{N}(0, R_i)$. Here we are using $\oplus$ to represent the composition operator and will use $\ominus$ to represent the inverse of a transformation such that $\ominus x \oplus x$ is the identity transformation.

Although Figure 4 shows just one constraint, in general we will want to integrate a set of constraints Z attached to different poses in the set $\overline{X}_i$ and this motivates the creation of a sum-of squares cost function

$$C(\mathbf{x}, Z) = \sum_{z_i \in Z}^{|Z|} \|z_i - h_i(x)\|^2_{\mathbf{R}_i} + \sum_{r=m+1}^{i} \|\hat{x_r} - x_r\|^2_{\mathbf{\Pi}_r} \quad (2)$$

where $||\mathbf{w}||^2_{\mathbf{A}} \triangleq \mathbf{w}^T \mathbf{A}^{-1} \mathbf{w}$ for some vector $\mathbf{w}$, and weighting matrix $\mathbf{A}$. For the sake of clarity, we are now writing (with a slight abuse of notation) the vertical concatenation of relative poses in $\overline{X}$ as the vector $\mathbf{x}$. Note that the first term in Equation 2 penalises a chain of transformations which, in concert, do not explain overhead constraints. The second term penalises deviation away from the raw pose-graph produced by the RBA subsystem. The cost $C(\mathbf{x}, Z)$ can be compactly expressed as a weighted inner product

$$\begin{aligned} C(\mathbf{x}, Z) &= \begin{bmatrix} \mathbf{z} - h(\mathbf{x}) \\ \hat{\mathbf{x}} - \mathbf{x} \end{bmatrix}^T \begin{bmatrix} \mathbf{R} & \\ & \mathbf{\Pi} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{z} - h(\mathbf{x}) \\ \hat{\mathbf{x}} - \mathbf{x} \end{bmatrix} \\ &= \left[\bar{\mathbf{z}} - g(\mathbf{x})\right]^T \mathbf{Q}^{-1} \left[\bar{\mathbf{z}} - g(\mathbf{x})\right] \end{aligned} \quad (3)$$

where $\mathbf{R}$ and $\mathbf{\Pi}$ are block diagonal matrices formed from individual $\mathbf{R}_i$ and $\mathbf{\Pi}_r$ respectively. Our task now is to minimize $C(\mathbf{x}, Z)$ with respect to $\mathbf{x}$. This is a standard problem and can be solved iteratively as $\mathbf{x}_{k+1} = \mathbf{x}_k + \delta\mathbf{x}$ where $\delta\mathbf{x}$ is a small change in $\mathbf{x}$. Iteration stops when $||\delta\mathbf{x}|| < \epsilon$ for some small number ϵ and $\delta\mathbf{x}$ is found by solving the linearized normal-equations

$$\mathbf{G}^T\mathbf{Q}^{-1}\mathbf{G}\delta\mathbf{x} = \mathbf{G}^T\mathbf{Q}^{-1}\left(\ \bar{\mathbf{z}} - g(\mathbf{x}_k)\ \right) \quad (4)$$

where

$$\mathbf{G} = \begin{bmatrix} \mathbf{H} \\ \mathbf{I} \end{bmatrix}$$

and $\mathbf{H}$ is the Jacobian of h,

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \\ \vdots \\ \mathbf{H}_{|Z|} \end{bmatrix}.$$

The i^{th} row of $\mathbf{H}$ is computed as

$$\mathbf{H_i} = \left[\frac{\partial h_i}{\partial x_m}, ..., \frac{\partial h_i}{\partial x_i}\right].$$

where each element $\frac{\partial h_i}{\partial x_p}$ has a simple form derived from application of the chain rule to Equation 1:

$$\frac{\partial h_i}{\partial x_p} = J_1\left({}^{m}x_p, {}^{p}x_i\right) J_2\left({}^{m}x_{p-1}, x_p\right) \quad (5)$$

and $J_1(x,y) = \frac{\partial x \oplus y}{\partial x}$ and $J_2(x,y) = \frac{\partial x \oplus y}{\partial y}$.

V. Results

We have applied the system to the road network around Oxford, England. Initial experiments demonstrate ~10 meter level accuracy, constant time operation, robustness to complex intersections, and robustness to false overhead road-correspondences. Figure 5 shows a ~1.1km sub-section from a ~100km dataset. Using overhead imagery of the road network,

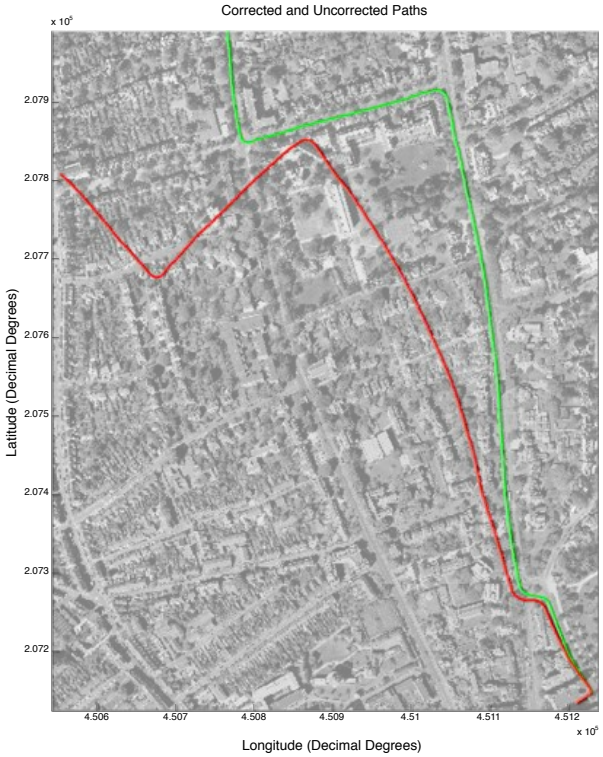

Figure 5: Part of the 100km Highway dataset RBA (red) and corrected onto the road (green).

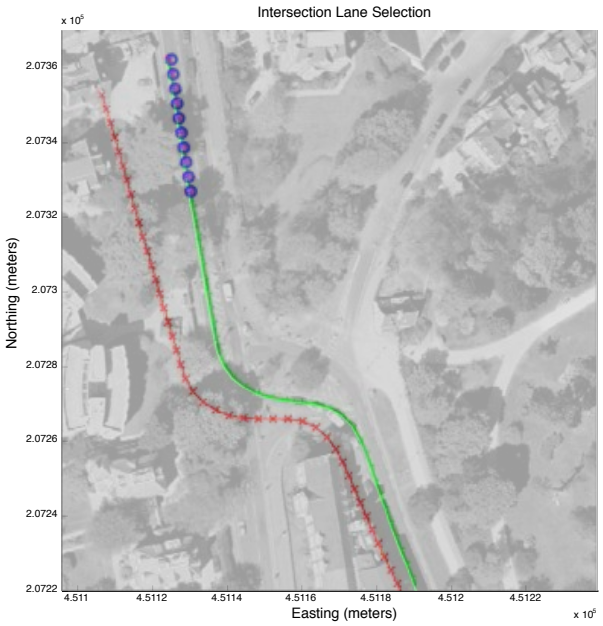

Figure 6: Weak pose-graph constraints in combination with strong relative SLAM constraints aid route selection at complex intersections. RBA (red), corrected (green), road correspondences (blue circles)

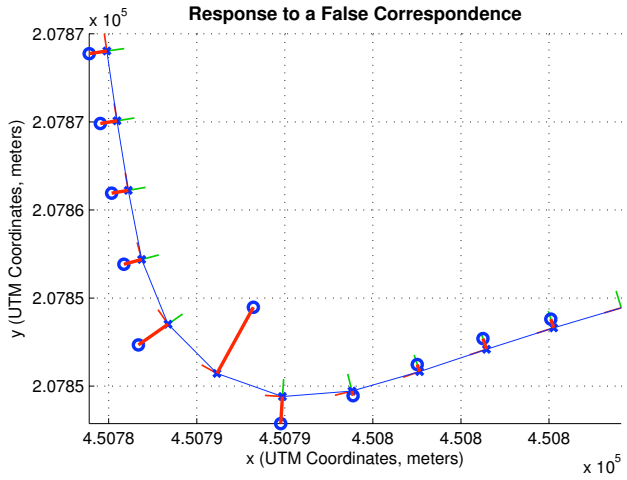

Figure 7: RBA accuracy overcomes false road correspondences. The vehicle poses are shown by the reg and gree axes

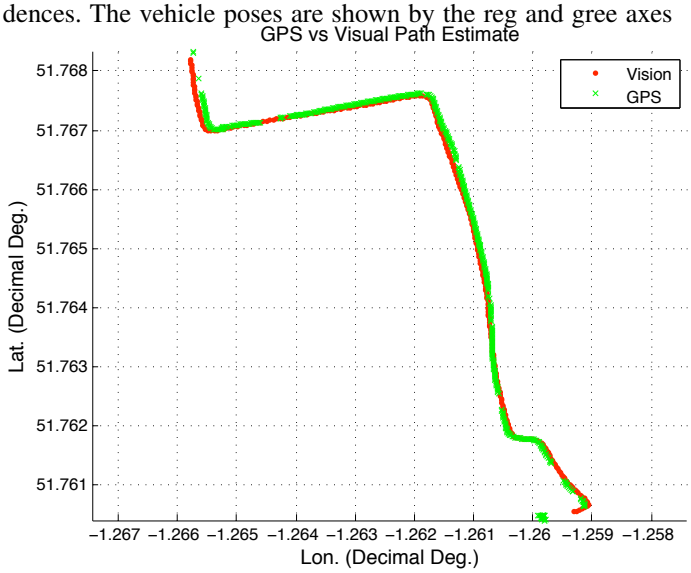

Figure 8: GPS vs visual path estimate for the start of the Highway data. Error in latitude and longitude are shown in Figures 9 and 10.

this demonstrates the ability to correct drift in relative SLAM estimates. Figure 6 shows the system navigating a potentially ambiguous portion of the road network (an intersection).

Figure 7 demonstrates robustness to false data association. Here one road correspondence is incorrect, but the rigidity of the local RBA path in conjunction with a majority correct correspondences, is (typically) sufficient to keep well localized.

Figures 8, 9 and 10 show estimated path vs. GPS. This demonstrates drift correction and ~10m accuracy.

VI. Discussion

The trajectory computed by the relative SLAM engine is locally accurate. This precision disambiguates potentially confusing situations, as incorrect matches have a relatively minor impact and are ignored (as in Figure 7). Here, in a difficult situation, the vehicle proceeds until it has driven past the ambiguity (an intersection), at which point data association recovers. Both RBA and our coarse-to-fine strategy

Figure 9: Error in Latitude relative to GPS (with a 10m CEP). An error of 1e-4 deg. equates to ~10m.

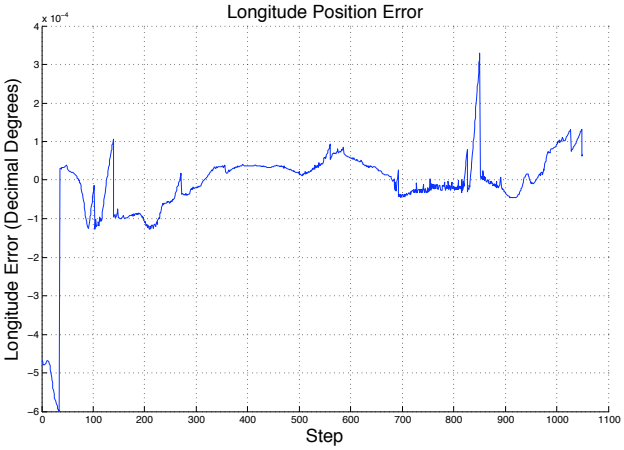

Figure 10: Error in Longitude relative to GPS (with a 10m CEP). An error of 1e-4 deg. equates to ~10m.

are designed to ensure a majority of good matches, so that the path can be swiftly adjusted onto the road. When the system is well localized by the coarse-to-fine strategy, the high-resolution pixel level matching is never grossly ambiguous. This makes it easy to reject false matches (due to shadows, change in road appearance, cars, weather, etc.) and only keep highly probable road matches.

Presently, the ~10m error we report is comparable with the error of our GPS unit. We believe the true performance is much better, so in the near future we would like to compare against RTK-GPS, which is purported to be accurate to ~2cm.

There are pathological and subtle cases we have not yet addressed. For instance, Figure 11 shows two difficult and interesting examples. In the first, a slight "Y-like" intersection leads to a false data association, and misleads the ensuing trajectory. In the second, a long section of straight road leads to poor position estimates in the direction of travel. The first problem is more difficult, though a multi-hypothesis tracker could follow both road-branches until the likelihood reduces to just one branch. The second problem will only occur when direct texture-to-texture matching fails - a situation we expect to be rare. In any case this concern is somewhat mitigated by our coarse to fine strategy and the use of synthesized super-resolution images ($\mathcal{I}^{+}$) .

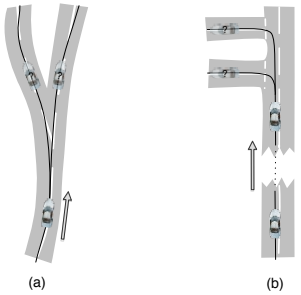

Figure 11: Future work will address ambiguity resolution for pathological cases, such as the "subtle y intersection" problem in (a), and the "unconstrained-travel-distance" problem in (b).

VII. Conclusion

We have explored the efficacy of combining overhead image-priors with a contemporary visual SLAM system and shown a working system in an urban environment. The principal contribution of this work is framework and implementation that demonstrates self-contained, infrastructure-free, bounded-error position estimation for a road vehicle. We are currently pursuing the application of the tool chain we described here to vast scales. Our intention is to extend this technique to alternative sensing modalities - in particular 2D and 3D laser - in no small part because of their suitability to inclement weather and low light conditions over and above that of vision.

VIII. Acknowledgments

The work reported in the paper is funded by an EPSRC DTA. We would also like to thank Christopher Mei for his help adapting his ESM image tracking software, which can be found at http://www.robots.ox.ac.uk/~cmei.

References

[1] Cooperative vehicle-infrastructure systems. http://www.cvisproject.org/en/cvis_project/.

[2] Cybernetic technologies for cars in the city. http://www.cybercars.org.

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

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

[5] R. Kummerle, B. Stederand C. Dornhege, A. Kleiner andG. Grisetti, and W. Burgard. Large scale graph-based slam using aerial images as prior information. In *Robotics Science and Systems*, 2008.

[6] J. Levinson, M. Montemerlo, and S. Thrun. Map-based precision vehicle localization in urban environments. June 2007.

[7] 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.

[8] 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.

[9] Gabe Sibley, Chris Mei, Ian Reid, and Paul Newman. Adaptive relative bundle adjustment. In *Robotics Science and Systems Conference*, Seattle, USA, June 2009.

[10] Sebastian Thrun and Michael Montemerlo. The graph slam algorithm with applications to large-scale mapping of urban structures. *Int. J. Rob. Res.*, 25(5-6):403–429, 2006.

[11] C. White, D. Bernstein, and A. Kornhauser. Some map matching algorithms for personal navigation assistants. *Transporta- tion Research Part C*, 8:91–108, 2000.