Generation and Exploitation of Local Orthographic Imagery for Road Vehicle Localisation

Ashley Napier and Paul Newman
Mobile Robotics Group University of Oxford
{ashley,pnewman}@robots.ox.ac.uk

***Abstract*—This paper is about road vehicle localisation based on vision using synthesised local orthographic imagery. We exploit state of the art stereo visual odometry (VO) on our survey vehicle to generate high precision synthetic orthographic images of the road surface as would be seen from overhead. The fidelity and detail of these images far exceeds that of aerial photographs. When undertaking subsequent passes of the same route, the vehicle is localised against the survey vehicle's trajectory by maximising the mutual information between the synthetic orthographic images and live image streams. Thus we explicitly leverage the gross appearance of the workspace rather than a discrete set of point features. We test our technique on data gathered from a road vehicle and show that centimeter-level precision is possible without the complexity and instability of contemporary feature based techniques.**

I. Introduction

Vehicle localisation has been a vigorously researched topic over the last few decades. For road vehicles especially, a popular recent approach is to use some combination of DGPS, inertial and 3D laser sensing coupled with a prior survey [2], [1]. Here, with an eye on road vehicle localisation, we investigate how we might achieve commensurate precision using just vision. During a survey stage we leverage a VO system to synthesise a continuous image strip of the road as seen from above, a synthetic local orthographic image. This strip need not be metrically correct over large scales (100m) but locally it provides an excellent template against which to match views obtained during subsequent traversals. In contrast to many registration techniques we do not attempt a feature based registration. Instead we seek the vehicle pose relative to the survey trajectory by maximising the mutual information between synthetic local orthographic images and the current view. The synthetic images allow localisation when traveling in either direction over the road surface. In some sense what we purpose here is a form of teach and repeat localisation, albeit one that negates the need for point features [5]. With this in hand global localisation is possible if the survey vehicle's trajectory has been post processed optimised into a single global frame [17]. This process of metric rectification is well studied [9], [6] and beyond the scope of this paper.

A. Background

The goal of this paper is high precision localisation without a reliance on external infrastructure or workspace modification. Localisation and pose estimation derived from local sensors suffers from compounding errors. For example stereo Visual Odometry (VO) produces locally metric maps and trajectories [7]. However, when extending to larger scales without correction the metric precision is lost and maps and trajectories become only topologically correct. Small angular errors, which over the course of a few hundred meters will lead pose estimates to be tens of meters in error. This makes 'knowing where you are' impossible without some sort of correction or reference to prior data.

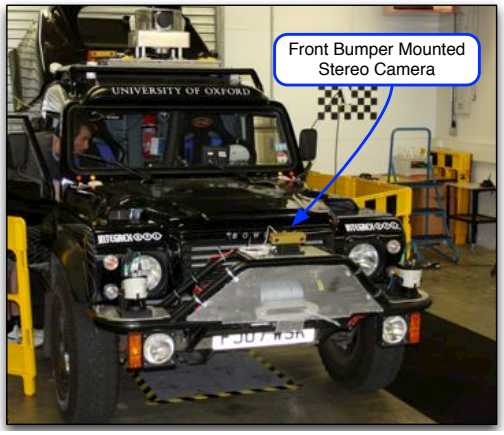

Figure 1. The Bowler Wildcat is MRGs latest mobile platform. It is a 4x4 all terrain vehicle supporting state of the art computing and sensors, including Stereo Vision used in this research.

B. Exploitation of Orthographic Imagery

Previous work using a VO system [12] attempted to correct for these small errors using aerial images as prior information, thereby maintaining the metric accuracy and global consistency of trajectory estimates. A coarse-to-fine approach was adopted where progressively finer refinements to the pose estimates were made by matching images from the local stereo camera to aerial images. This approach produced pose estimates commensurate with the performance of off-the-shelf GPS over kilometer scales. Pink et al[13], [14] used a similar approach but instead extracted road markings from aerial and camera images to perform the localisation. Another approach by Kummerle et al[8] extracted edges of buildings form aerial images for corrections using a 2D laser based system.

The suitability of aerial images for reliable and accurate correction of VO poses for road vehicles however, has its limitations. The road surface is often occluded by trees and

bridges and image resolution is of the order of tens of centimeters per pixel.

C. Our Approach

In this work we replace the aerial images with synthetic local orthographic images of the road surface. These images are generated by a survey vehicle and vehicle localisation for subsequent traversals of a route is done relative to the survey vehicles trajectory. This approach allows generation of orthographic images under bridges, trees and in man made structures such as multi story car parks. This method also has the advantage that only things which can be seen from a road vehicles perspective are included in the images, excluding roof tops and grassy fields etc. Levinson et al[10] adopted a similar approach using laser reflectance maps, however their approach relied heavily on GPS and IMU and was based in a global frame.

Here we represent the orthographic image relative to a survey vehicle trajectory which is not necessarily metrically correct over large scales. Our synthetic orthographic image generation is therefore not tied to a global frame, so does not require metric global consistency[16]. We also have no reliance on GPS, any external infrastructure or workspace modification. Synthetic orthographic images are generated at a resolution two orders of magnitude higher than the best available aerial images (5mm per pixel). Subsequent traversals of surveyed routes by a follow vehicles can then be localised against the survey trajectory using these high resolution, high fidelity orthographic images.

II. Motivation

One may ask why bother generating orthographic imagery? We have a VO system that extracts point features (in our case SIFT features [11]) and saves them into a map (see Figure 3) why not localise using those?

Feature based methods have been shown to be very sensitive to relatively small changes in view point, leading to a significant drop off in matched features available for localisation, as made explicit in [5]. Figure 2 demonstrates how localisation using a feature based approach is not sufficient. Only 40% of the features are matched against the previous trajectory and after only $70m$ with a small deviation from the previous traversal there are not enough matched features for localisation, leading to irrecoverable failure. In the interests of further illustrating this point, 2 manual relocalisations were preformed each of which eventually led to localisation failure.

Many point features are ephemeral, including vegetation, parked vehicles, and road speckle, which are all unlikely to be seen on subsequent traversals of the same route. The gross appearance of the road surface however, does not change for view point shifts experienced by a road vehicle, even when traveling in the opposite direction. This is because the road surface is a largely planar, man-mad structure, we successfully exploit both of these properties in this work.

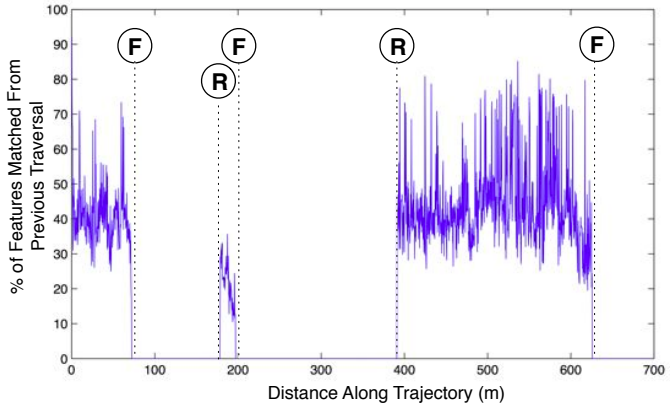

Figure 2. Shows the percentage of features matched against a previous trajectory of 700 meters. F indicates a failure i.e. not enough features matched to achieve localisation and R denotes a manual relocalisation. It can be seen that only $\sim 40\%$ of features are matched and the system fails 3 times over a 700m trajectory requiring manual intervention.

III. Survey Trajectory

Let us define the notation and terms on which we will rely. A vehicle trajectory $\mathcal{T}$, as generated by the VO, is a set of relative $\mathbb{SE}3$ transformations ${}^{i}x_{i+1}$ between consecutive vehicle poses X_i . The relative transform between X_i and X_{i+1} is written as ${}^{i}x_{i+1}$. In this way $\mathcal{T} = \{{}^{0}x_1, {}^{1}x_2, \cdots, {}^{N-1}x_N\}$ is a trajectory of N vehicle poses (see Figure 3). Note that in the interest of clarity we shall sometimes refer to X_i as a vehicle node rather than as a pose, the reason being that one can imagine $\mathcal{T}$ as a chain graph of nodes linked by relative transformations. Each node in the graph has one or more 3D landmark features attached to it. The jth landmark attached to vehicle node i is denoted as ${}^{i}l_j$. We use the function $L(X_i)$ to represent the act of producing a local metric scene $\mathbf{M_i}$ around X_i which contains a set of local vehicle nodes and all attached landmarks in the frame of X_i (so X_i becomes the origin).

$$\mathbf{M_i} = \left\{ \begin{array}{c} X_{i-1}, X_i, X_{i+1}, X_{i+2} \\ {}^{i-1}l_1, {}^{i-1}l_j, .., {}^{i}l_0, {}^{i+1}l_{1+j}, ..., {}^{i+2}l_1, ... \end{array} \right\} \quad (1)$$

$$X_{i+2} = X_i \oplus {}^{i}x_{i+1} \oplus {}^{i+1}x_{i+2} \quad (2)$$

$$X_{i-1} = X_i \ominus {}^{i-1}x_i \quad (3)$$

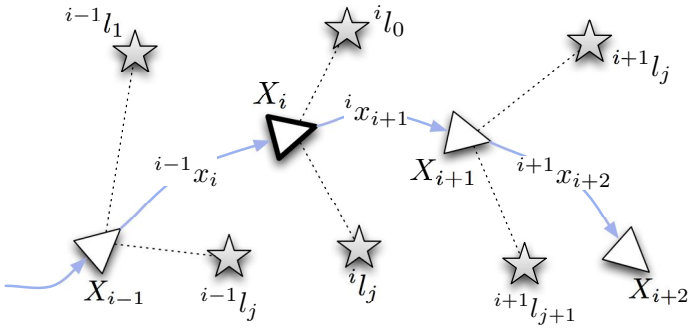

Figure 3. Notation for a simple VO trajectory and scene $\mathbf{M_i}$. Relative poses between frames are represented by arrows and lowercase ${}^{i-1}x_i$, Vehicle poses relative to X_i are represented by triangles and uppercase X_{i-1}. The location of the jth observed landmark from pose X_i are represented by stars ${}^{i}l_j$.

Where $\oplus$ and $\ominus$ represent the composition and inverse composition operators respectively, such that ${}^{i}x_{i+1} \ominus {}^{i}x_{i+1}$ is the identity transformation. In Section IV we shall describe how $\mathbf{M_i}$ can be used to generate a synthetic local orthographic image around X_i.

IV. Generating Local Orthographic Images

At run time we will use knowledge gleaned from a previously recorded excursion (survey) to ensure we stay localised relative to the prior survey. A fundamental competency on which we depend is the availability/generation of a synthetic orthographic image around a particular node X_i in the survey trajectory $\mathcal{T}$. We use the function $\mathcal{I}_\pi(X_i)$ to denote the generation of this image which utilises the local metric scene $\mathbf{M}_i$ defined above. $\mathcal{I}_\pi(X_i)$ can be computed at run time, but can also be precomputed. The process begins with the extraction of a ground plane using the landmarks ${}^{i}l_j$ in $\mathbf{M}_i$ and RANSAC [4] to solve.

$$\boldsymbol{b}_{\boldsymbol{i}}^{\boldsymbol{\pi}}.\hat{\boldsymbol{n}}_{\boldsymbol{i}}^{\boldsymbol{\pi}} = \boldsymbol{l}.\hat{\boldsymbol{n}}_{\boldsymbol{i}}^{\boldsymbol{\pi}} \tag{4}$$

Where $\boldsymbol{b}_{\boldsymbol{i}}^{\boldsymbol{\pi}}$ is the base, $\hat{\boldsymbol{n}}_{\boldsymbol{i}}^{\boldsymbol{\pi}}$ the normal and $\boldsymbol{l}$ an arbitrary point on the plane. Note that in our experimental set up, the front bumper stereo camera (Figure. 1) is orientated such that at least half of the view contains road surface, assuming we are traversing relatively smooth urban roads. This orientation ensures many landmark measurements correspond to points on the road surface aiding ground plane estimation. Based on the stereo camera's orientation a Region Of Interest (ROI) $\boldsymbol{\mathcal{V}}^{\boldsymbol{\mathcal{I}}} = [\mathcal{V}_1^{\mathcal{I}}, ..., \mathcal{V}_J^{\mathcal{I}}]$ in the left camera image is set as a region likely to only contain road surface. Here $\mathcal{V}_j^{\mathcal{I}} = [u, v, 1]^T$ is a pixel location in homogeneous co-ordinates of the jth vertex of the ROI. The intersection of the rays associated with each $\mathcal{V}_j^{\mathcal{I}}$ and the local ground plane ($\boldsymbol{b}_{\boldsymbol{i}}^{\boldsymbol{\pi}}$, $\hat{\boldsymbol{n}}_{\boldsymbol{i}}^{\boldsymbol{\pi}}$) are calculated. $\boldsymbol{\mathcal{V}}_{\boldsymbol{i}}^{\boldsymbol{\pi}} = [\mathcal{V}_{i,1}^{\pi}, ..., \mathcal{V}_{i,J}^{\pi}]$ is then the ROI projected onto the local ground plane around pose X_i.

$$\mathcal{V}_{i,j}^{\pi} = \lambda_j K^{-1} \mathcal{V}_j^{\mathcal{I}} \tag{5}$$

$$\lambda_j = \frac{\boldsymbol{b}_{\boldsymbol{i}}^{\boldsymbol{\pi}}.\hat{\boldsymbol{n}}_{\boldsymbol{i}}^{\boldsymbol{\pi}}}{\left[K^{-1}\mathcal{V}_j^{\mathcal{I}}\right].\hat{\boldsymbol{n}}_{\boldsymbol{i}}^{\boldsymbol{\pi}}} \tag{6}$$

where K is the matrix of camera intrinsics, $K^{-1}\mathcal{V}_j^{\mathcal{I}}$ is the ray associated with $\mathcal{V}_j^{\mathcal{I}}$ and λ_j is the distance along $K^{-1}\mathcal{V}_j^{\mathcal{I}}$ to the intersection with the ground plane. A homography $\mathcal{H}_i$ is then generated from $\boldsymbol{\mathcal{V}}^{\boldsymbol{\mathcal{I}}}$ and $\boldsymbol{\mathcal{V}}_{\boldsymbol{i}}^{\boldsymbol{\pi}}$ such that

$$\boldsymbol{\mathcal{V}}^{\boldsymbol{\mathcal{I}}} = \mathcal{H}_i \boldsymbol{\mathcal{V}}_{\boldsymbol{i}}^{\boldsymbol{\pi}} \tag{7}$$

$\mathcal{H}_i$ is then used to project the texture in the survey images ROI taken at X_i into an orthographic image which we call $\mathcal{I}_\pi(X_i)$. We do this for all poses in scene the $\mathbf{M}_i$. The camera frame rate of $20Hz$ coupled with the survey vehicle's velocity of approx. 20kph leads to adequate overlap between consecutive $\boldsymbol{\mathcal{V}}_{\boldsymbol{i}}^{\boldsymbol{\pi}}$ (road regions of interest projected onto the orthographic image $\mathcal{I}_\pi(X_i)$). This presents an opportunity to combine ROIs for consecutive poses in $\mathbf{M}_i$ by taking an average of intensity values. This generates an image of length

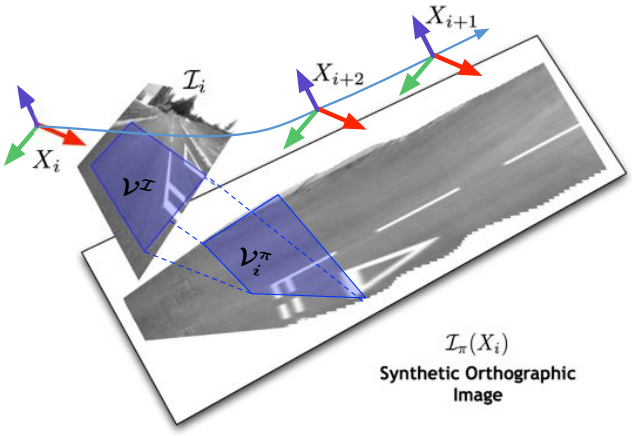

Figure 4. Shows the view $\mathcal{I}_i$ at pose X_i from the survey trajectory being projected onto the local ground plane to produce an orthographic image $\mathcal{I}_\pi(X_i)$. Road regions of interest $\boldsymbol{\mathcal{V}}^{\boldsymbol{\mathcal{I}}}$ and $\boldsymbol{\mathcal{V}}_{\boldsymbol{i}}^{\boldsymbol{\pi}}$ shown in blue.

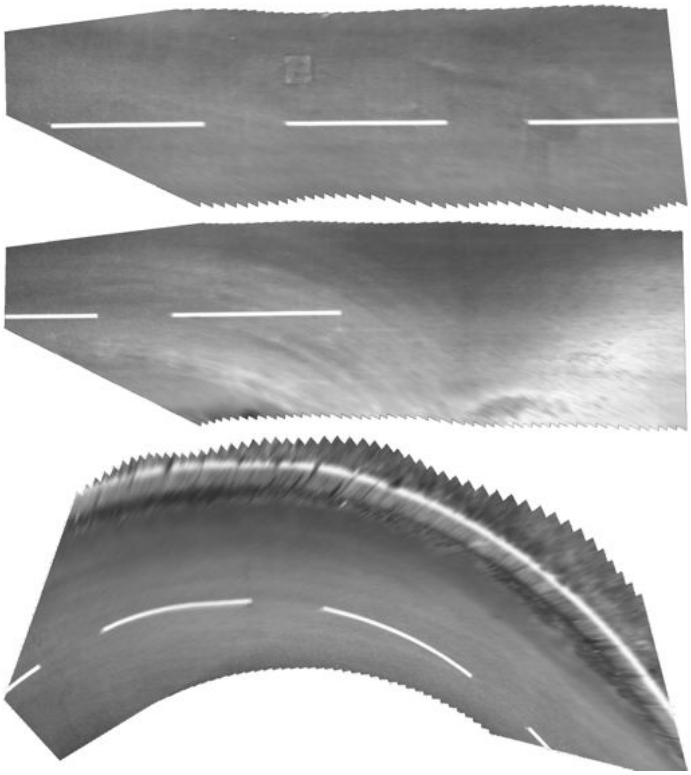

Figure 5. 10 meter segments of the synthetic local orthographic images generated from the 700m test route. The top two are from straight segments and the third is part of a 90^o corner.

defined by the poses in $\mathbf{M}_i$ of the road surface as seen from overhead in the vicinity of X_i (Figure 4). Results presented in this paper use a resolution of 5mm per pixel in $\mathcal{I}_\pi(X_i)$ (Note that this resolution was chosen as a balance between accuracy and storage requirements, approx. 10MB/km). Examples of images $\mathcal{I}_\pi(X_i)$ associated with our test route are shown in Figure. 5. The image alignment accuracy is of the order $10^{-3}m$ and the images contain a high level of detail.

V. Localisation

Consider now an image $\mathcal{I}_k$ acquired on a subsequent traversal of a surveyed route at time k in the vicinity of X_i. The pose of the vehicle can now be represented relative to X_i with ${}^{i}t_k$ which is the transformation between X_i and the location of the vehicle at time k (See Figure. 6). If required the global pose of the vehicle X_k is then simply.

$$X_k = X_i \oplus {}^{i}t_k \tag{8}$$

At run time our Stereo VO system provides a continual stream of estimates of the relative pose between camera frames v_k. In the absence of any other knowledge we could use this to infer our trajectory open loop. If however we can leverage the synthetic orthographic images to correct relative poses from the VO we would be in a position to track our motion (stay localised) relative to the survey trajectory. Furthermore, if as a new stereo pair is presented to us, we could use v_k to seed a guess for the transformation ${}^{i}t_{ko}$ between the new camera frame and the survey trajectory, for the example in Figure. 6 ${}^{i}t_{ko}$ would be expressed as.

$${}^{i}t_{ko} = \ominus {}^{i-1}x_i \ominus {}^{i-2}x_{i-1} \oplus {}^{i-2}t_{k-1} \oplus v_k \tag{9}$$

Of course as we move we need to track our location relative to sequential poses in the trajectory - X_i will change as we move. However this is a trivial data association problem; we can simply predict the transition to a new reference pose using v_k as indicated by our VO system. The goal of this paper is then to develop a way to hone this initial estimate ${}^{i}t_{ko}$ and we shall do so by comparing the live view $\mathcal{I}_k$ and that predicted by a hypothosised view of the synthetic orthographic image $\mathcal{I}_\pi(X_i)$ at ${}^{i}t_k$.

We shall pose the task of finding ${}^{i}t_k$ as an optimisation problem. The objective function used is based on Mutual Information (MI), which was originally defined in [15]. If we knew ${}^{i}t_k$ perfectly then the projection $proj(\mathcal{I}_\pi(X_i), {}^{i}t_k)$ (hypothesised view) of the road lying in $\mathcal{I}_\pi(X_i)$ into the live view $\mathcal{I}_k$ would overlap completely. Conversely, if the pose is in error, the two views will not align and in particular will exhibit a markedly reduced amount of MI. The optimisation therefore finds a relative pose ${}^{i}\hat{t}_k$ which maximises image alignment by maximising MI [18] (see Fig. 7). Note for normal operation we use images from the left stereo camera.

A. Mutual Information for Image Alignment

We use Mutual Information rather than a simple correlation based approach as it has shown to be robust against varying lighting conditions and occlusions [3]. The MI between two images $\mathcal{I}$ and $\mathcal{I}^*$, intuitively can be thought of the information shared between the two images. It is defined as follows.

$$MI(\mathcal{I}, \mathcal{I}^*) = H(\mathcal{I}) + H(\mathcal{I}^*) - H(\mathcal{I}, \mathcal{I}^*) \tag{10}$$

The MI is obtained by evaluating the Shannon entropy of the images individually $H(\mathcal{I})$ and $H(\mathcal{I}^*)$, and then evaluating the joint entropy $H(\mathcal{I}, \mathcal{I}^*)$. The entropy of a single image is a measure of how much information is contained within the image.

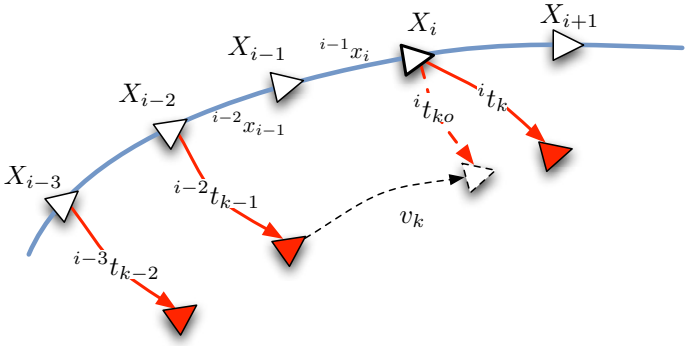

Figure 6. Localisation Framework - The survey trajectory is shown in blue with vehicle nodes/poses X_i. A subsequent traversal of the route localised against the survey trajectory is shown in red ${}^{i}t_k$, the current relative VO pose v_k and subsequent corresponding seed for localisation ${}^{i}t_{ko}$ are indicated by dashed lines.

$$H(\mathcal{I}) = -\sum_{n=0}^{N} p_{\mathcal{I}}(n) log(p_{\mathcal{I}}(n)) \tag{11}$$

Where $p_{\mathcal{I}}(n)$ is the probability of a pixel in image $\mathcal{I}$ having intensity n. An image can therefore be thought of as a random variable with each pixel location x having a distribution defined by , $p_{\mathcal{I}}(n) = p(I(x) = n)$ for $n \in [0, N]$, where N is the maximum intensity (in our case 255, as we are using 8-bit grayscale images). The joint entropy is defined by

$$H(\mathcal{I}, \mathcal{I}^*) = -\sum_{n=0}^{N}\sum_{m=0}^{N} p_{\mathcal{I}\mathcal{I}^*}(n, m) log(p_{\mathcal{I}\mathcal{I}^*}(n, m)) \tag{12}$$

Where $p_{\mathcal{I}\mathcal{I}^*}(n, m) = p(\mathcal{I}(x) = n, \mathcal{I}^*(x) = m)$ the joint probability of intensity co-occurrances in both images. The MI can then be written as

$$MI(\mathcal{I}, \mathcal{I}^*) = \sum_{n=0}^{N}\sum_{m=0}^{N} p_{\mathcal{I}\mathcal{I}^*}(n, m) log(\frac{p_{\mathcal{I}\mathcal{I}^*}(n, m)}{p_{\mathcal{I}}(n) p_{\mathcal{I}^*}(m)}) \tag{13}$$

As an implementation detail it was found empirically that evaluating the MI over all possible pixel intensity values had little advantage over histogramming intensitites into bins. Quantising the intensity values into 16 bins has a welcome smoothing effect on the cost surface and eases optimisation. Another advantage of using MI over other plausible measures such as SSD (Sum of Square Distances) is that it is meaningfully bounded. The minimum MI is zero and the maximum

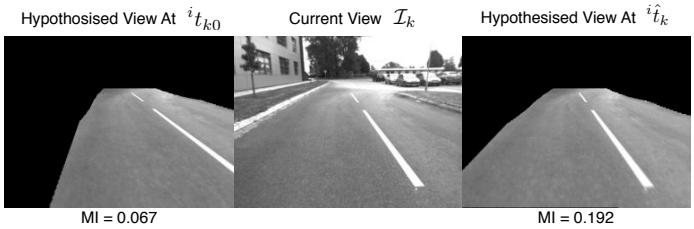

Figure 7. Example of view alignment between the live view and hypothosised views and their corresponding values for mutual information. Left shows the hypothosised view at ${}^{i}t_{ko}$ provided by v_k, Middle shows the live image $\mathcal{I}_k$ and Right shows the hypothosised view at the corrected pose ${}^{i}\hat{t}_k$.

is the minimum value of information contained within each of the images $min(H(\mathcal{I}),H(\mathcal{I}^*))$. The maximum possible information for an image is also bounded as an image with a uniform distribution of pixel values.

The problem of estimating our current pose relative to a pose X_i in the survey trajectory $\mathcal{T}$ then reduces to solving

$$ {}^{i}\hat{t}_k = \underset{{}^{i}t_k}{argmax} \left\{ MI(\mathcal{I}_k, proj(\mathcal{I}_\pi(X_i), {}^{i}t_k)) \right\} \tag{14} $$

VI. Implementation & Results

As the Stereo VO has high metric accuracy over small distances $\sim 10m$, the deviation from the initial position estimates ${}^{i}t_{ko}$ are relatively small, of the order of centimeters. Two approaches have been implemented for estimating ${}^{i}t_k$ the first is a fast approximate $\mathbb{SE}2$ and the second a full $\mathbb{SE}3$ pose correction.

A. $\mathbb{SE}2$ *Pose Corrections*

Our application domain is road vehicles so in our first approach to reduce complexity and increase speed we confine ${}^{i}t_k$ to in road plane motion, reducing our search space to $\mathbb{SE}2$. We still however maintain the $\mathbb{SE}3$ pose information which allows us to correct for rolling and pitching during cornering and accelerations respectively. Rather than solve eq 14 iteratively by for example using non-linear Gauss Newton methods, we take advantage of the small search radius and use a histogram filter to evaluate an approximation to eq. 14.

The in plane motion approximation has the consequence of reducing the sensitivity of the matching step to high frequency image content, such as fine texture on the tarmac. Very small errors in pitch, roll or height cause misalignments which stop the matching process from leveraging this fine detail. As we generally operate in urban environments where the vast majority of roads have distinct road markings this was deemed to be an acceptable trade off for speed. However, for short periods where there are no road markings the histogram filter can fall into local minima. This can lead to errors in our estimations of ${}^{i}\hat{t}_k$ which has the effect of pulling the trajectory off course. In order to avoid this, we first we compute the difference between the corrected pose ${}^{i}\hat{t}_k$ in the survey trajectory and the initialisation from the Stereo VO ${}^{i}t_{ko}$.

$$ e = {}^{i}\hat{t}_k \ominus {}^{i}t_{ko} \tag{15} $$

If e is greater than a threshold we invoke the right camera, perform localisation on $\mathcal{I}_{k,right}$ and check for consensus, if the pose estimates from both $\mathcal{I}_{k,right}$ and $\mathcal{I}_k$ are commensurate we adopt the pose correction into the trajectory. If the pose estimates don't agree we ignore the match and simply adopt ${}^{i}t_{ko}$ into the trajectory, we repeat this until matching can be reestablished. In essence when the image matching step fails the system falls back to raw VO and runs in Open Loop mode. Here there is an interesting area of future work, how might we learn e and what constitutes as agreement beyond obvious statistical tests.

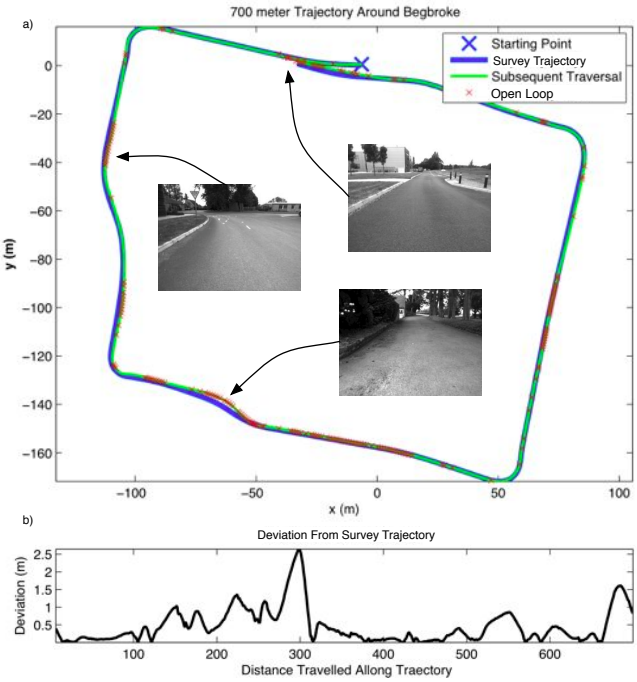

Figure 8. (a) Route around the begbroke site, approximately 700m loop approximated $\mathbb{SE}2$ corrections, (b) localised pose distances from survey trajectory.

To evaluate the performance of our algorithm we conducted experiments on data collected from our autonomous vehicle platform the 'WildCat' (see Figure 1). This modified Bowler WildCat is equipped with a multitude of sensors and 32 processor cores available for computation. For this work we use the front mounted Point Gray Bumblebee2, running at 640x480 at 20 fps and only 9 cores. The Stereo VO runs at frame rate on a single core. The localisation runs on 8 cores at a lower frequency, approximately $1Hz$, as correction is not required for every VO pose. Currently the algorithm is implemented in Matlab, however the use of a histogram filter lends itself perfectly to a GPU implementation, currently in development, which promises significant speed up and reduction in resource usage.

Figure 8.(a) shows the 700m survey trajectory (blue) overlaid with a subsequent traversal localised against the survey in (green). It should be noted that although the subsequent traversal is commensurate with the survey trajectory they should not align perfectly. This is because the survey vehicle may have driven at different positions on the road. The pose distance from the survey trajectory should therefore vary, but be bounded as shown in Figure 8.(b). Figure 8.(a), also shows when the consistency check is active and the system runs in open loop, it can be seen that this occurs when there are no road markings present. The local metric accuracy of the Stereo VO system allows localisation to run in open loop mode until the matching process is successfully reestablished. For the route shown here, matching was available 70% of the time and the vehicle completed a full 700m loop of the site without any intervention or manual relocalisation as was required with feature based matching, Figure 2.

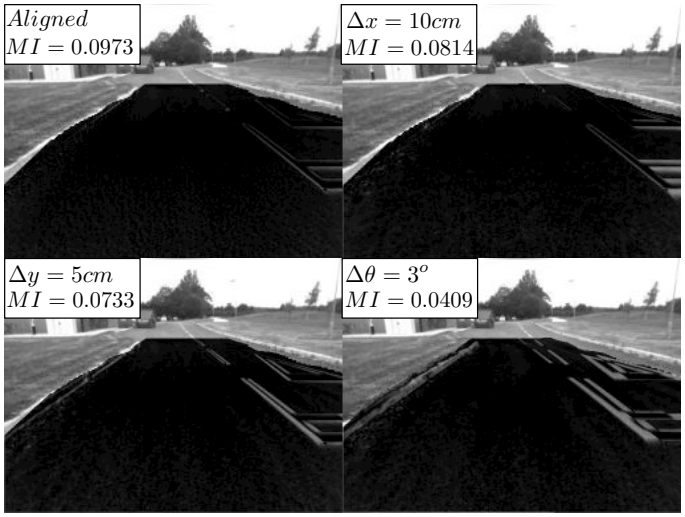

Figure 9. Top left shows the difference in image intensities between a driver view $\mathcal{I}_k$ correctly localised relative to the survey trajectory and the hypothosised view $proj(\mathcal{I}_\pi(X_i),^i t_k)$ at $^i t_k$ (Note white pixels correspond to larger differences in intensity). Top right to bottom right show the effect of artificially perturbing the pose by $\Delta x = 10cm, \Delta y = 5cm, \Delta\theta = 5^0$ respectively. It can be seen there is a significant miss alignment and corresponding drop in Mutual Information.

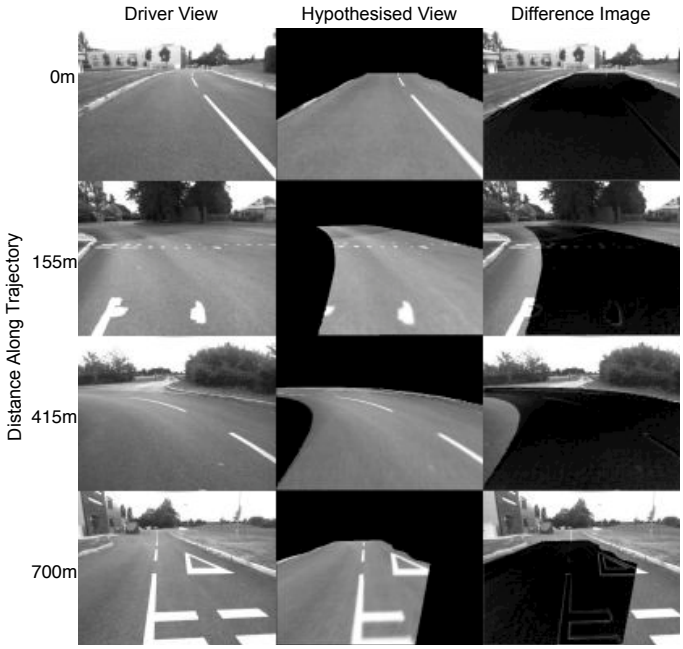

Figure 10. Driver views $\mathcal{I}_k$ and associated hypothosised views $proj(\mathcal{I}_\pi(X_i),^i \hat{t}_k)$ demonstrating cm level alignment at various points along the 700m trajectory with reference to Figure 9.

B. Demonstrating Centimeter Precision

As we don't have centimeter accurate ground truth we demonstrate the precision of the localisation by comparing localised images $\mathcal{I}_k$ and the corresponding artificially perturbed hypothosised views $proj(\mathcal{I}_\pi(X_i),^i t_k \oplus \epsilon)$, where ϵ is a perturbation of the order of centimeters. Figure 9, shows that by artificially perturbing the vehicle pose from its true value we can see significant image misalignment and a corresponding drop in MI. Figure 10 shows several frames taken from a subsequent traversal of the survey route localised using our method. From the difference images it can be seen that the two views are aligned to centimeter precision. We also demonstrate convexity in Figure 11, here the objective function (MI) is plotted around a true pose for all 6 degrees of freedom.

C. Full $\mathbb{SE}3$ Pose Corrections

Solving for all six degrees of freedom with no in plane motion constraint allows the matching step to leverage much finer detail and high frequency image content such as texture in the tarmac. Figure 11 demonstrates how a significant peak and convexity is maintained in the Objective Function (MI) at the correct pose under various conditions. Here we can see that the Mutual Information as a correlation measure is robust against partial road occlusions, varying lighting and weather conditions.

With the current MATLAB implementation, solving for $\mathbb{SE}3$ pose corrections is not real time. However, we are currently developing an OpenCL implementation which has demonstrated two orders of magnitude speed up. Solving for the full $\mathbb{SE}3$ pose corrections will allow the localisation method to be used in less urban environments, will reduce the amount of time in Open Loop mode as well as the need for consensus checking as the matching process is more readily available on a wider variety of road surfaces.

VII. Conclusion

This work presents a methodology for generating and exploiting synthetic local orthographic images to achieve centimeter precision road vehicle localisation without any infrastructure or work space modification. The method improves accuracy by an order of magnitude on a previous method using off the shelf aerial images [12]. We use Stereo VO to generate synthetic orthographic images from a survey vehicle which are far superior in terms of resolution and fidelity to available aerial images. Our approach also allows us to generate orthographic images in areas unavailable to aerial photography such as under bridges, trees and covered areas. These images provide an high fidelity and stable template for view matching as unlike feature based systems we use the gross appearance of the road surface ahead of the vehicle. Our approach avoids all the tracking and data association required by feature based approaches. We demonstrate centimeter level accurate localisation and pose tracking is demonstrated on a 700m trajectory around our campus as well as robustness to partial occlusion and varying weather and lighting conditions.

VIII. Acknowledgments

The work in this paper is funded by an EPSRC DTA, EPSRC Leadership Fellowship EP/I005021/1 supporting Paul Newman and used the WildCat research platform provided by BAE Systems. The authors would also like to thank Dr. Gabe Sibley and Mr. Winston Churchill for their assistance with the Stereo VO system.

References

[1] R Behringer, W Travis, R Daily, D Bevly, W Kubinger, W Herzner, and V Fehlberg. RASCAL - an autonomous ground vehicle for desert driving

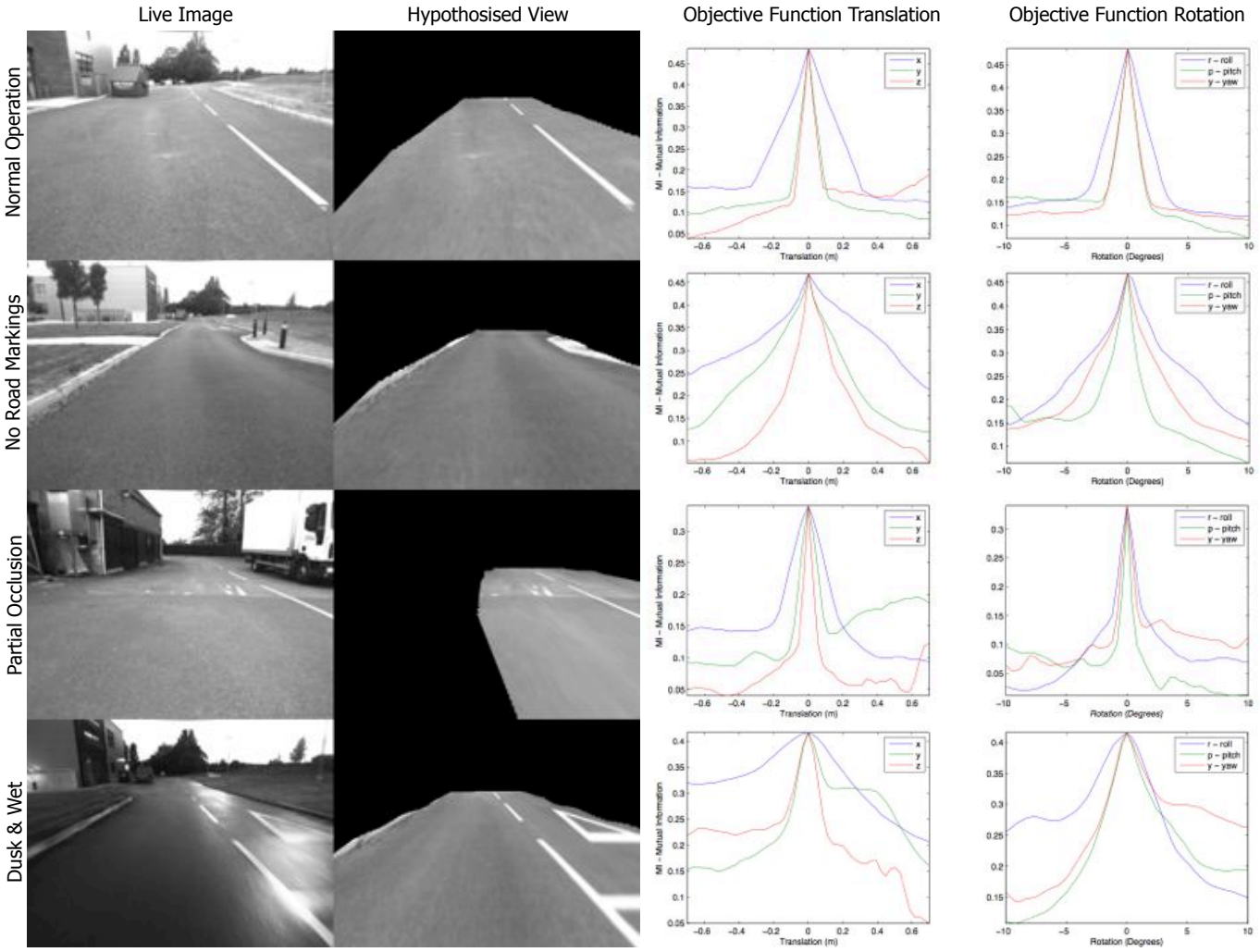

Figure 11. The objective function (MI) demonstrating convexity and robustness to partial occlusion, weather and lighting conditions. The pose of the vehicle it shifted from it true value, the corresponding drop in MI demonstrates convexity in all 6 degrees of freedom plotted at $1cm$ and 0.2^o intervals respectively. In our experiments we found the average correction to be $2cm$ and $< 1^o$ so well within the basin of attraction. The histogram filter also naturally finds distinct peaks.

in the DARPA Grand Challenge 2005. In *Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE*, pages 644–649, 2005.

[2] Q Chen, U Ozguner, and K Redmill. Ohio State University at the 2004 DARPA Grand Challenge: developing a completely autonomous vehicle. *Intelligent Systems, IEEE*, 19(5):8–11, 2004.

[3] Amaury Dame and Eric Marchand. Mutual Information-Based Visual Servoing. *Robotics, IEEE Transactions on*, (99):1–12, 2011.

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

[5] Paul Furgale and Timothy D. Barfoot. Visual teach and repeat for long-range rover autonomy. *Journal of Field Robotics*, 27(5):534–560, 2010.

[6] Wolfram Burgard Giorgio Grisetti, Cyrill Stachniss. Non-linear constraint network optimization for efficient map learning. *IEEE Transactions on Intelligent Transportation Systems*, 10:428–439, 2009.

[7] K. Konolige, M. Agrawal, and Joan Solà. Large scale visual odometry for rough terrain. In *Proc. International Symposium on Research in Robotics (ISRR)*, November 2007.

[8] R Kümmerle, B Steder, C Dornhege, and A Kleiner. Large scale graph-based SLAM using aerial images as prior information. *Proc. of Robotics: Science and Systems*, 2009.

[9] Rainer Kümmerle, Giorgio Grisetti, Hauke Strasdat, Kurt Konolige, and Wolfram Burgard. g2o: A General Framework for Graph Optimization. *Proceedings of the International Conference on Robotics & Automation (ICRA) 2011*, pages 1–7, February 2011.

[10] J Levinson, M Montemerlo, and S Thrun. Map-based precision vehicle localization in urban environments. *Proceedings of the Robotics: Science and Systems*, 2010.

[11] D Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.

[12] A Napier, G Sibley, and P Newman. Real-time bounded-error pose estimation for road vehicles using vision. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 1141–1146, 2010.

[13] O Pink and F Moosmann. Visual features for vehicle localization and ego-motion estimation. *Intelligent Vehicles Symposium, IEEE*, pages 254–260, 2009.

[14] O Pink and C Stiller. Automated Map Generation from Aerial Images for Precise Vehicle Localization . *Inteligent Transport Systems Converence (ITSC) 2010*, pages 1517–1522, August 2010.

[15] CE Shannon. The mathematical theory of communication. *Bell system technical journal*, 27:379–423, 1948.

[16] G Sibley, C Mei, I Reid, and P Newman. Vast-scale Outdoor Navigation Using Adaptive Relative Bundle Adjustment. *The International Journal of Robotics Research*, 2010.

[17] S Thrun. The graph SLAM algorithm with applications to large-scale mapping of urban structures. *The International Journal of Robotics Research*, 2006.

[18] William M Wells III, Paul Viola, Hideki Atsumi, Shin Nakajima, and Ron Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1(1):35–51, March 1996.