Outdoor SLAM using Visual Appearance and Laser Ranging

P. Newman, D. Cole and K. Ho
Oxford University Robotics Research Group. Email: {pnewman,klh,dmc}@robots.ox.ac.uk

***Abstract*— This paper describes a 3D SLAM system using information from an actuated laser scanner and camera installed on a mobile robot.The laser samples the local geometry of the environment and is used to incrementally build a 3D point-cloud map of the workspace. Sequences of images from the camera are used to detect loop closure events (without reference to the internal estimates of vehicle location) using a novel appearance-based retrieval system. The loop closure detection is robust to repetitive visual structure and provides a probabilistic measure of confidence. The images suggesting loop closure are then further processed with their corresponding local laser scans to yield putative Euclidean image-image transformations. We show how naive application of this transformation to effect the loop closure can lead to catastrophic linearization errors and go on to describe a way in which gross, pre-loop closing errors can be successfully annulled. We demonstrate our system working in a challenging, outdoor setting containing substantial loops and beguiling, gently curving traversals. The results are overlaid on an aerial image to provide a ground truth comparison with the estimated map. The paper concludes with an extension into the multi-robot domain in which 3D maps resulting from distinct SLAM sessions (no common reference frame) are combined without recourse to mutual observation.**

I. Introduction

We would like to have a robot perform Simultaneous Localization and Mapping (SLAM) outdoors. We aim to replace the now ubiquitous 2D, planar maps generated from moving indoors on flat floor with rich 3D maps built from vehicles moving on more general outdoor terrain in which estimated vehicle trajectories are now embedded in $\mathbb{R}^6$. Moreover we wish to do this over large areas and accommodate large loops. We do not constrain ourselves to try and do this using only one sensor, for example laser or camera. Instead this work brings together two complementary threads of research - SLAM using a 3D laser scanner and loop closure detection using photometric information. The resulting system which we describe here is, to our knowledge, the first time a 3D laser-vision SLAM system with automated loop closure detection has been showcased working in a typical outdoor urban environment.

We begin in Section II by describing the 3D laser scanner and the scan-match based SLAM framework we choose to employ. In Section III we introduce the sequential, appearance based loop-closure detection algorithm. We then describe how, following the detection of a potential loop closure, a rigid transformation is produced that relates the current location of the vehicle to a previous one. In Section IV we describe how this transformation can be fused with current map and trajectory estimates despite the presence of potentially gross errors in vehicle location and effect the loop closure. Section V presents results generated by this system when applied to a data set of an outdoor site and superimposes the results on an aerial site photograph for comparison. In Section V-A the work is extended to show how the visual loop closing system can be used to guide the fusion of multiple maps (for example built simultaneously by a different vehicle or the same vehicle on a different day) with no common co-ordinate frames to yield a larger combined map. The paper concludes with a discussion of current shortcomings and further work.

II. Geometry from Laser

Previous work on 3D SLAM on mobile platforms includes work by Davison and Kita [5], who use purely vision, and more recent work by Weingarten and Siegwart [24] who use 3D laser range data. They subsequently extract planar features from the scans and maintain parameterized estimates in a probabilistic manner using the SP formulation [2]. The work by Surmann, Nuchter, Lingemann and Hertzberg in [22] and [23] is notable. They use sequential registrations to fuse multiple 3D laser range scans together, and when loop-closing, use batch update techniques to redistribute registration error. This work presents strong results, but it is not clear how loop closures are detected. The 3D SLAM system we use is a 3D extension of the delayed state formulation in [1] and [17] and has much in common with work in [8], [14] and more recently [7]. The underlying SLAM representation is a state vector of past vehicle poses. At a suitable interval the state vector is augmented with a new vehicle pose using odometry information $\mathbf{u}(k)$ between times k and $k+1$:

$$\mathbf{x}(k+1|k) = \begin{bmatrix} \mathbf{x}(k|k) \\ \mathbf{x}_{vn}(k|k) \oplus \mathbf{u}(k+1) \end{bmatrix} \quad (1)$$

$$= \begin{bmatrix} \mathbf{x}_{v1}^T & \dots & \mathbf{x}_{vn}^T & \mathbf{x}_{vn+1}^T \end{bmatrix}^T (k+1|k) \quad (2)$$

where $\oplus$ is the $\mathcal{SE}_3$ transformation composition operator and $\mathbf{x}_{vi}$ is the i^{th} vehicle pose in the state vector. Here we are using the standard conditional $(p|q)$ notation to denote an estimate at time p conditioned on observations up until time q. Associated with this state vector is a covariance matrix $\mathbf{P}$ with the following structure:

$$\mathbf{P}(k+1|k) = \begin{bmatrix} \mathbf{P}(k|k) & \mathbf{P}_{vp}(k+1|k) \\ \mathbf{P}_{vp}(k+1|k)^T & \mathbf{P}_v(k+1|k) \end{bmatrix} \quad (3)$$

where $\mathbf{P}_v$ is the covariance of the newly added vehicle state. As the vehicle moves, a 'nodding' laser scanner returns a stream of planar scans at different elevations. Each pose

$\mathbf{x}_v(k)$ in the state vector has a cloud of points 'attached' and referenced to it which we will refer to as a scan $\mathbf{S}_k$.

The decision on how to segment the laser observation stream and hence decide when to augment the state-vector is a function of vehicle control and distance travelled, preferring to cluster scans from linear motion over scans built while the vehicle executes substantial rotations. This is described in [4].

A 3D registration procedure can be applied to any two scans $\mathbf{S}_i$ and $\mathbf{S}_j$ to yield an observation $\mathbf{T}_{i,j}$ of the rigid transformation between poses $\mathbf{x}_v(i)$ and $\mathbf{x}_v(j)$ in the state vector. This observation is applied to the state vector using the usual data fusion machinery (in this case we use the standard EKF equations). Under normal conditions, i and j are sequential, in which case the $\mathbf{u}$ in equation 1 can be replaced with $\mathbf{T}_{i,j}$. However during loop closing the poses being related by $\mathbf{T}_{i,j}$ stem from the vision-based sub-system described in section III, and will be temporally very different ($j >> i$). In fact, it is likely they will also be spatially far apart because of accumulated errors around large loops — the consequences of which are considered in Section IV. The evolution of the state vector and its uncertainty is illustrated in Figure 1 which was generated using an outdoor data set which shall be used throughout this paper. The figure shows the evolution of the state's individual poses, from our vehicle's initial position, until a just before a loop closure was detected visually .

What remains to be discussed is precisely how the occurrence of a loop closure can be detected. A simple and sometimes effective approach would be to look to the SLAM p.d.f to detect loop closure as in [16]. A sufficiently small Mahalanobis distance between two poses could then indicate loop closure. However, this assumes that the estimate of the underlying joint p.d.f. is *not* in gross error. Unfortunately there *are* gross errors in position estimate — especially after traversal around long loops. This is why loop closure detection is so difficult.

Consider, for example, the situation depicted in Figure 1, in which the combined uncertainties around current and prior vehicle locations come nowhere near suggesting loop closure (even though the vehicle has actually returned to the origin). One might comment that something has gone drastically wrong with the estimator to yield such a poor estimate of position. It is true to say that this data set was chosen because it produced a particularly spectacular gross error in trajectory estimate. Nevertheless, we maintain that whichever SLAM estimator is used, however good the odometry is or whatever onboard inertial sensors are employed, a data set could be generated over *some* terrain or scale that results in gross errors in both map and trajectory estimates.

In this work we make no recourse to the p.d.f. estimate to detect loop closures. Instead we substitute the technique discussed in Section III, which examines time sequences of images and finds similarities in appearance between the recent past and image sequences taken in the distant past. The algorithm is then able to return a transformation estimate between the two poses concerned. This becomes the initializing solution to the iterative laser scan registration described above and in [4]. Once found, all that remains is to actually 'close the loop'. Section IV describes how this is achieved.

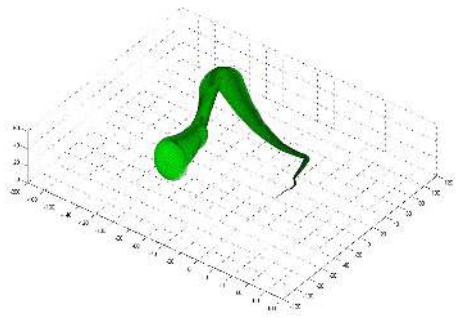

Fig. 1. The evolution of the state vector's individual poses from the vehicle's initial position, until the moment before the loop closer prompted a loop closure between the last (current) and first pose. The corresponding 1σ marginal 'x, y, z' uncertainty ellipsoids are plotted around each individual pose state. Note the vehicle has actually returned to the origin — far outside the possibilities admitted by the uncertainty ellipses. The grid is marked in intervals of 20m.

III. Detecting Loop Closure With Vision

In this section we will describe a new approach to detecting loop closure using images. The method is appearance based and uses the similarity between local scene descriptions to find statistically significant pairings between sequences of images. The images we are considering here are ordered — the vehicle captures them sequentially as it moves through its workspace meaning images captured close in time also have a spatial proximity.

We make the reasonable assumption that if two images I_u and I_v look the same then there is an elevated probability that they are indeed two views of the same place. If I_u was captured at time step k_u and I_v at k_v then the vehicle's poses at times k_u and k_v should be close[1]. However, finding just one image pair correspondence may not always be a strong enough basis on which to suggest a loop-closure event. As a simple illustrative example, the exterior of buildings frequently present repetitive architectural structure, and indoors many doorways look the same. However if we can chain correspondence pairs together to form *sequences* of paired images we can increase our confidence that the two strands are multiple views of the same scene — they are similar throughout an extended spatial area.

The method described in this section explicitly tackles difficulties in differentiating scenes due to repetitious low-level descriptors (common texture) or broad, background inter-scene similarity (common large-scale features[2]). The procedure produces sequences of paired, time stamped images. Each sequence represents a putative loop closure, and has an associated probability, conditioned on the totality of collected images, that this is a genuine loop closure.

[1] assuming that not all of the scene content is in the far field

[2] like windows, plants or Victorian architecture — see Figure 4

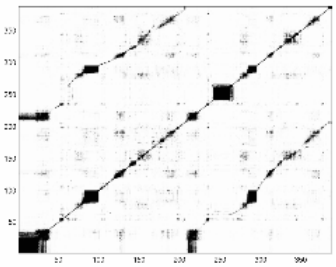

Fig. 2. The above illustrates a typical visual similarity matrix with loop closure appearing as dark off-diagonal streaks.

A. Computing Similarity

A camera mounted on the vehicle captures the local scene (left and right) every few meters of traversal. Each image is passed into a processing pipeline which first extracts affine invariant regions of interest — we wish to be able to detect these regions from varied view points — and then parameterizes them into a suitable descriptor. We typically use the SIFT descriptor [13] and the Harris Affine Detector [15] but our method works with any detector-descriptor pairing that offers wide base line stability and suitably rich descriptions. Each image I_u is transformed into a set of descriptors $\{d_1 \cdots d_n\}$, where, when using SIFT descriptors, each d_i is a $128D$ vector. The number of descriptors n, will in general be different for each image. As more and more images are acquired the total number of descriptors keeps increasing. As suggested in [20] clustering the accumulating descriptor set yields a visual vocabulary providing "visual words" which collectively describe the images. We apply a simple agglomerative, "leader-follower" clustering algorithm which yields a vocabulary $\mathcal{V}$, size $\mid \mathcal{V} \mid$, of visual words $\{\hat{\mathbf{d}}_\mathbf{1}, \hat{\mathbf{d}}_\mathbf{2}...\}$. We note that we cannot assume that in the context of the whole vocabulary, all words have equal descriptive value. Some words may apply in practically every image while others in just a few. The inverse document frequency weighting scheme of [11] is an established way of assigning a weight $w_i = \log N/n_i$ to an index term ($\hat{\mathbf{d}}_\mathbf{i}$) as a function of the number of documents (images) in which it appears (n_i) and the total number of documents (images) N. We are now able to describe each image I_u as a vector of weights $\mathbf{I}_u = [u_1 \cdots u_{|\mathcal{V}|}]^T$ where

$$u_i = \begin{cases} w_i & \text{if for } \hat{\mathbf{d}}_\mathbf{i} \in I_u \\ 0 & \text{otherwise.} \end{cases}$$

A central requirement of our technique is to be able to quantify the similarity between any two scenes u and v which we denote as $S(u, v)$. The cosine distance is a suitable measure so that

$$S(u,v) = \frac{\sum_{i=0}^{|\mathcal{V}|} u_i v_i}{\sqrt{\sum_{i=0}^{|\mathcal{V}|} u_i^2}\sqrt{\sum_{i=0}^{|\mathcal{V}|} v_i^2}}. \quad (4)$$

This similarity function [3] allows the creation of a *Similarity Matrix* $\mathtt{M}$ which encodes the similarity between all N images. Each element $\mathtt{M}_{i,j}$ is $S(i, j)$ the similarity score between $\mathbf{I}_i$ and $\mathbf{I}_j$. A typical Similarity Matrix (and one that will be processed extensively as a example throughout this paper) is shown in Figure 2. Two things are immediately apparent: firstly the strong diagonal which stems from all images being self similar, and secondly the presence of off-diagonal streaks. These are the loop closures we are seeking to detect — sequences of temporally separated, yet similar scenes.

B. Sequence Extraction

As the vehicle moves through its work space it creates a sequence of images $\mathcal{I} = [I_1, I_2 \cdots]$. We pose the loop closure detection problem as finding two subsequences of $\mathcal{I}$, $\mathcal{A} = [a_1, a_2 \cdots]$ and $\mathcal{B} = [b_1, b_2...]$ where a_i and b_i are index variables, whose overall similarity strongly suggests that the vehicle is revisiting a region.

We use the notation $u \Leftrightarrow v$ to denote the pairing on grounds of similarity between images I_u and I_v. Importantly, there is nothing to say that $a_i \Leftrightarrow b_j$ should imply $a_{i+1} \Leftrightarrow b_{j+1}$. It could be that image a_{i+1} matches image b_j as well, perhaps implying that two sequential images in $\mathcal{I}$ are identical because the vehicle has stopped or, more troubling, is imaging a scene with a repetitive structure (which we shall come to soon). We now describe a modified form of the Smith-Waterman algorithm [21], a dynamic programming algorithm which we use to find $\mathcal{A}$ and $\mathcal{B}$. A matrix H is constructed in which element, $H_{i,j}$, is the maximal cumulative similarity score of a sequence of pairs of images ending with pairing $\langle I_i, I_j \rangle$. Within S, three move types are possible: diagonal, horizontal and vertical. The latter two, although viable, are less preferable (they cause one-to-many matching) and so have a penalty term δ associated with them. Hence depending on whether the move is from $H_{i-1,j-1}, H_{i,j-1}$ or $H_{i-1,j}$, $H_{i,j}$ becomes

$$H_{i,j} = \begin{cases} H_{i-1,j-1} + M_{i,j} & H_{i-1,j-1} \text{ maximal,} \\ H_{i,j-1} + M_{i,j} - \delta & H_{i,j-1} \text{ maximal,} \\ H_{i-1,j} + M_{i,j} - \delta & H_{i-1,j} \text{ maximal} \\ \alpha \max(H_{i-1,j-1}, H_{i,j-1} H_{i-1,j}) & M_{i,j} \leq \tau \text{ , } H_{i,j} > \tau \\ 0 & \text{otherwise} \end{cases}$$

The fourth case, where $0 < \alpha < 1$, allows for gaps in the sequence of matches typically caused by an obscured field of view (e.g. a pedestrian walking in front of the camera). The constant τ is a tolerance threshold which can be set using the statistical analysis in section III-D. When every cell has been visited the maximally scoring sequence $\langle \mathcal{A}, \mathcal{B} \rangle$ of paired images can be unwound by back-tracing through H starting at the maximum element in H whose value we denote as $\eta_{\mathcal{A},\mathcal{B}}$.

C. Removing Common-Mode Similarity

The procedure described thus far works well in environments with few visually ambiguous or repetitive scenes (see

[3]Implementation note : because the majority of the elements in the vectors $\mathbf{I}_u$ and $\mathbf{I}_v$ will be zero (they generally use only a fraction of the entire vocabulary), the calculation in 4 benefits greatly from an efficient implementation that iterates only over non-zero members.

Figure 2). We now ask how the algorithm will perform in an environment that is more visually confusing, such as the one resulting in the visual similarity matrix shown on the left hand side of Figure 3. An off-diagonal dark line starts at around image 400 — this is the start of the genuine loop closure. However there are also numerous dark (mutually similar) off-diagonal regions. These are typically caused by repetitive imaging of architectural features like windows, long brick walls or broadly homogenous foliage. We loosely describe these as "themes". Themes cause collections of words to appear *together* across multiple scenes resulting in a (correct) degree of mutual similarity. The resulting blocks and stripes in $\mathtt{M}$ make reliable sequence extraction problematic. We wish to remove the *effects* of these themes in the similarity matrix so that the sequence detection relies more on the unusual similarities between scenes than the common-mode terms. We achieve this by a rank reduction technique which we now describe.

The similarity matrix is decomposed into a sum of rank one matrices formed from outer products

$$M = \sum_{i=1}^{N} \mathbf{v}_i \lambda_i \mathbf{v}_i^T \tag{5}$$

where λ_i is the i^{th} eigenvalue of $\mathtt{M}$ and $\mathbf{v}_i$ the corresponding eigenvector. The left hand column of Figure 4 shows the first three outer products for the $\mathtt{M}$ in the L.H.S of Figure 3. The Matrix $\mathtt{M}$ is a real symmetric matrix and so each $\mathbf{v}_i \lambda_i \mathbf{v}_i^T$ is a rank one approximation to $\mathtt{M}$. If a theme is responsible for the dominant structure in $\mathtt{M}$ then, because $\sum_{i=1}^{r} \mathbf{v}_i \lambda_i \mathbf{v}_i^T$ is the best rank-r approximation to $\mathtt{M}$ under the Frobenius norm, we should expect its effect in $\mathtt{M}$ to be captured in the dominant eigenvalues / vectors. Thus, we can diminish the effect of visual ambiguity / repetitive scene structure by reconstructing $\mathtt{M}$ by omitting the first r terms of the summation in 5. We shall now discuss how to choose r.

For an $n \times n$ $\mathtt{M}$, we define the relative significance, $\rho(r)$ of λ_r as

$$\rho(r) = \lambda_r / \sum_{k=r}^{n} \lambda_k \tag{6}$$

Using this we can measure the complexity of decomposition of $\mathtt{M}$ as an entropy

$$H(\mathtt{M}, r) = -\sum_{k=r}^{n} \rho(r) log(\rho(r)). \tag{7}$$

The case that $H(\mathtt{M}) = 0$ corresponds to an ordered and redundant $\mathtt{M}$ which can be represented by a single eigenvector. $H(\mathtt{M}) = 1$ corresponds to a similarity matrix where all eigenvectors are equally expressive capturing equally important structure. Hence we are motivated to sequentially remove outer-products from $\mathtt{M}$ until $H(\mathtt{M})$ is maximised leaving a similarity matrix in which no one single theme dominates. We may replace $\mathtt{M}$ with a rank reduced version

$$\mathtt{M}' = \sum_{i=r^\star}^{N} \mathbf{v}_i \lambda_i \mathbf{v}_i^T \quad r^\star = \arg\max_r H(\mathtt{M}, r) \tag{8}$$

The typical effects of this procedure are illustrated in Figure 3 The left hand column of Figure 4 shows the first three rank-1 matrices removed when rebuilding $\mathtt{M}$ according to Equation 8 and the typical images responsible for them are shown on the right. Note that this is a soft association used for illustration only and is derived by looking at which cells in the removed $\mathbf{v}_i \mathbf{v}_i^T$ are maximal and that the off diagonals in Figure 3 have been thickened to make them visible in this p.d.f document.

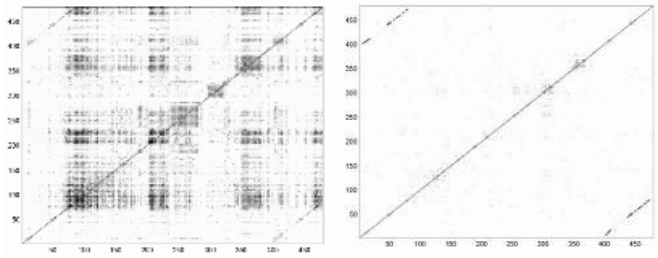

Fig. 3. The left hand figure shows a similarity matrix constructed from images collected while executing the trajectory shown in Figure 9. The right hand figure shows the structure of the "cleaned" similarity matrix after the rank reduction procedure described in Section III-C.

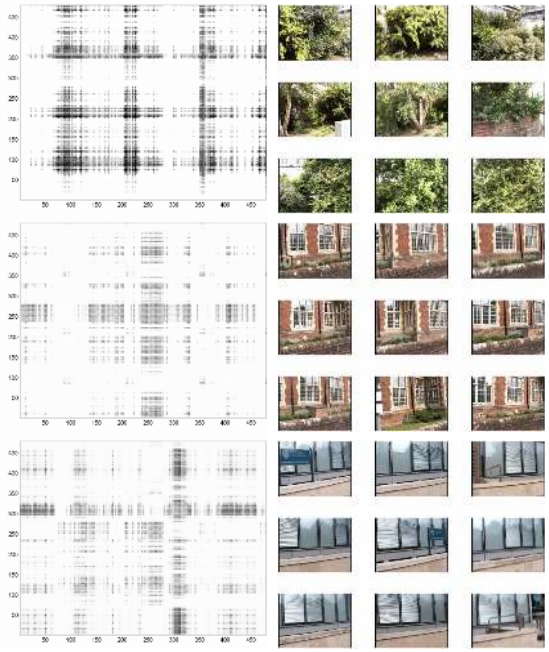

Fig. 4. The first three (of four) rank one matrices removed from the $\mathtt{M}$ shown in Figure 3 and the nine images most responsible for their dominance.

D. Sequence Significance

Having performed the rank reduction described above, the sequence extraction can be performed as before with the

effects repetitive visual structure now removed. The last requirement of our loop closure detection scheme is to assign some meaningful significance to the maximal sequence $\langle \mathcal{A}, \mathcal{B} \rangle$. It could, after all, be only marginally better than a randomly choosen route through $\mathtt{M}$. To achieve this we use the extreme value distribution (E.V.D.) to model the distribution of the maximum alignment score $\eta_{\mathcal{A},\mathcal{B}}$. The rows and columns of $\mathtt{M}$ are randomly shuffled and for each perturbation, a new maximum alignment score calculated. The E.V.D. describes the distribution of $\eta_{\mathcal{A},\mathcal{B}}$ as

$$p(\eta_{\mathcal{A},\mathcal{B}}) = \frac{1}{\beta} \exp^{-z} \exp^{-\exp^{-z}}, \quad z = \frac{\eta_{\mathcal{A},\mathcal{B}} - \mu}{\beta} \tag{9}$$

We proceed by estimating the mode and width parameters, μ and β, by fitting (we use a Levenburg-Marquadt scheme) to a histogram of the Montecarlo-generated alignment scores. Equipped with estimates $\hat{\mu}$ and $\hat{\beta}$ and the closed form C.D.F of the E.V.D. we can evaluate the probability of scores greater than or equal to $\eta_{\mathcal{A},\mathcal{B}}$ conditioned on all N images:

$$P(\eta \geq \eta_{\mathcal{A},\mathcal{B}} \mid \mathtt{M}) = 1 - \exp^{-\exp^{z}} \tag{10}$$

Equation 10 allows the evaluation of the probability that an extracted sequence of image matches $\langle \mathcal{A}, \mathcal{B} \rangle$ with score $\eta_{\mathcal{A},\mathcal{B}}$ could have been generated at random from $\mathtt{M}$. The differences between the sequence score $\eta_{\mathcal{A},\mathcal{B}}$ obtained from the original, temporally ordered $\mathtt{M}$ and those obtained from the randomly shuffled versions are solely attributable to the topology or connectedness of the spatial locations at which the vehicle captured the images. Thus Equation 10 can be used to evaluate the probability, conditioned on all previous scene appearances, that the detected sequence does indeed indicate a bona-fide loop closure.

E. How Many Loops?

An important distinction between global localisation and loop-closing is that in the former it is often known *a-priori* that a correspondence between a vehicle's local scene and a stored representation of the workspace exists. In the case of loop-closing this is not the case — the vehicle may never revisit the same location. It is also possible that within the totality of images, $\mathcal{I}$, multiple loop closure events are captured. The probabilistic formulation in Section III-D allows for both these situations. After rank reduction and then distribution fitting, sequences are extracted from $\mathtt{M}$ in decreasing order of alignment score, $\eta_{\mathcal{A},\mathcal{B}}$, until the probability of false positives associated with $\eta_{\mathcal{A},\mathcal{B}}$ becomes excessive. We typically set a threshold of 0.5%.

F. Estimating the Loop Closure Geometry

The sub-system described in the previous section produces a set of images and times (from the time-stamps of the images) which suggest the occurrence of loop closure. In other words we have a strong suspicion that the vehicle is near to a previously known location. One option would be to use the times to index into the state vector (which is after all a sequence of past poses) to find which previous pose i occupied the scene we are now revisiting at time j. However this approach has problems when it comes to undertaking a laser scan match to deduce a precise estimate of the interpose transformation $\mathbf{T}_{i,j}$. Without a reliable prior or "seed solution" the iterative scan matching method we adopt [4] frequently converges to an incorrect minima. At the same time exhaustive search in 6D is prohibitively slow. In this work we use the putative loop closure image sequences $\langle \mathcal{A}, \mathcal{B} \rangle$ and 3D laser data to estimate $\mathbf{T}_{i,j}$ – the later being used to remove scale ambiguity.

Consider the following common projective model of two identical cameras with projection matrices P and P' [9]. A homogenous 3D image scene point $X = [X, Y, Z, 1]^T$ imaged at $x = PX$ for the first camera and $x' = P'X$ for the second camera. Without loss of generality the origin can be fixed at the center of the first camera and if the second camera center is parameterized by a rotation matrix R and a translation t with respect to the origin then P and P' can be written $K[I|0]$ and $K[R|t]$ respectively, where K is the matrix of intrinsic camera parameters. In the case of calibrated cameras (K known) the image points, x and x', are related by the "Essential matrix" E such that $x'^T E x = 0$. This is a linear constraint on the elements of E which can be obtained by solving the equivalent constraint $\widetilde{x}^T \widetilde{E} = 0$ where $\widetilde{x}$ is a linear combination of the elements of x and x', and $\widetilde{E}$ is the vector of elements of E.

The determination of relative pose between camera positions by decomposing an essential matrix has been used to good effect in robot localization [12], [19] and SLAM navigation [6]. Given two image views of the same scene, five point of correspondences are selected for use in an implementation of the "five point algorithm " described in [18] which, notably, is capable of dealing with coplanar correspondence points.

Given two views of the same scene, five or more point of correspondences (this whole procedure should run inside a RANSAC routine) are selected and by stacking the vector $\widetilde{x}^T$ for all five points, a 5 x 9 matrix is obtained allowing four solutions for the nine elements of $\widetilde{E}$ (and thus E) to be obtained. The matrix E has a particularly convenient structure. It can be written in terms of R and t as $[t]_x R$ where $[t]_\times$ denotes the 3×3 skew symmetric (cross product) matrix constructed from t. Given the elements of E this decomposition yields four possible solutions for R and t up to scale. The correct solution is selected by application of suitable constraints on the world points X.

Because we know the rigid transformation between the laser scanner and the camera, 3D laser data can be expressed in the camera frame and projected onto the imaging plane (see Figure 5). The range of the correspondence points used to deduce E can now be estimated from neighboring laser range points. This immediately leads to the removal of the scale ambiguity in t.

Given estimates of R and t, the iterative laser scan matching can proceed with these estimates as an initial solution to $\mathbf{T}_{i,j}$. We note that R and t are not perfect owing to inaccuracies in K and imprecise knowledge of the instantaneous laser to camera transformation and the fact that the centers of the

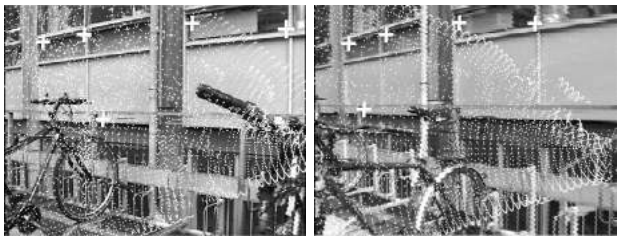

Fig. 5. Calculating the loop closure geometry. Using five or more correspondences (crosses) between Harris Affine features, the transformation between two views can be deduced up to scale. 3D laser data is back projected onto the images (white dots) to find the range of the interest points and thus remove the scale ambiguity. The distribution of laser points is due to the combination of vehicle motion and "nodding" motion of the laser scanner.

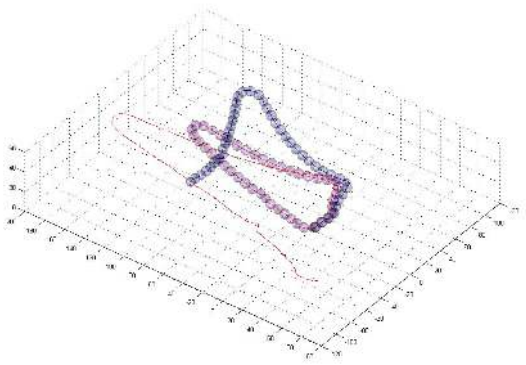

Fig. 6. Vehicle poses prior to any loop closure are shown as a thin line. Small spheres indicate rigid groups of poses whose inter-frame transformations are optimised. The open loop is the result of attempting to apply a lop closure constraint in one step while the closed loop results from iterative non-linear optimisation.

interest points may not coincide exactly with back-projected laser data. Nevertheless, this is a sufficiently accurate relative pose estimation to enable the scan matching to converge and fine tune the transformation estimate.

There remains a finite, albeit small, possibility that the loop closure indicated so far is in fact a false positive. It appears, conditioned on all images, to be statistically significant, while R and t describe the views of the scene with tolerable back-projection error. One could resort to the estimates of vehicle location to glean an idea of the credibility although how one could define "credible" in the presence of gross p.d.f errors is an open question. As an alternative we can use the quality of the final scan match to make the final accept/reject decision. This is described in the companion paper [4].

IV. Enforcing Loop Closure

Once a loop closure with a corresponding loop closing transformation $T_{i,j}$ is found, it is used to update the state vector. The system described in Section II might be expected to redistribute (as dictated by the state covariance matrix $\mathbf{P}$) any gross errors around the circuit. This is exactly what would happen if the update equations involved were linear - however in this case $T_{i,j}$ expresses a non-linear constraint on the state vector.

Figure 6 shows the effect of naively enforcing the loop closure constraint on the pre-loop closing state vector using the EKF update equations. The result is a discontinuous 'x, y, z' trajectory with highly unrealistic orientations.

This is a well known problem, caused by linearization which is considered in detail in [3]. Motivated by this, we side-step the problem by performing constrained non-linear optimisation around the loop, essentially performing multiple, incremental and iterative changes to the map instead of a single application of a constraint. Prior to loop closing the state vector is a set of n stacked $\mathcal{SE}_3$ poses referenced to the origin. However simple application of reference frame transformations yields an equivalent, open chain of n sequential pose to pose transformations, $\mathbf{T}_{i,i+1} \forall i \in [0:n]$, around the loop. Assuming, without loss of generality, that loop closure is detected between pose n and pose 1 a final $(n+1)^{th}$ transformation $\mathbf{T}_{n,1}$ can be added to the set which closes the chain of transformations. We can also obtain a measure of the uncertainty in each $\mathbf{T}_{i,i+1}$ in the form of a covariance matrix $\Sigma_{i,i+1}$ by recalling the results from the original scan-matching procedure and the nominal inter-pose odometry. Given the set of initial set of relative transformations $\mathcal{T} = [\mathbf{T}_{1,2}, \mathbf{T}_{2,3} \cdots \mathbf{T}_{n,1}]$ and associated uncertainties $\Sigma = [\Sigma_{1,2}, \Sigma_{2,3} \cdots \Sigma_{n,1}]$ we set ourselves the task of finding some new set of inter-pose $\mathcal{SE}_3$ transformations $\mathcal{T}^\star$ such that

$$\mathbf{C}(\mathcal{T}^\star) = \sum_{i=1}^{n+1} (T_i^\star - T_i)^T \Sigma_n^{-1} (T_i^\star - T_i) \tag{11}$$

is minimised subject to the loop closure constraint

$$T_1 \oplus T_2 \oplus \cdots T_n \oplus T_{n+1} = \mathbf{0} \tag{12}$$

where, with a slight abuse of notation, T_i and $T_i^\star$ are the i^{th} elements of sets $\mathcal{T}$ and $\mathcal{T}^\star$ respectively and $\oplus$ is the $\mathcal{SE}_3$ composition operator.

The procedure just described allows each inter-pose transformation to be adjusted following loop closure. We have experimented with reducing the degrees of freedom in the optimisation by arbitrarily segmenting the state vector into rigid sub-maps containing some small number, $m > 1$, of vehicle poses. The optimisation then occurs over the reduced space of inter-submap transformations. While this is of course to the detriment of overall map quality we have subjectively found its effect to be small and worthwhile compared to the increase in loop-closing speed.

V. Results

The complete system we have described in this paper has been successfully applied to several data sets, one of which is described in detail in this section. Figure 9 shows the final estimated map and trajectory of the vehicle as it traverses the perimeter of a cluster of buildings. The inset to the figure shows a detail of part of the overall picture. The "stripping" on the floor is an artifact of the nodding motion of the laser scanner. The final estimated vehicle trajectory is superimposed upon an aerial photograph of the workspace. Additionally a metric grid has been placed over the area of interest. The

TABLE I

Driven Path	370.95m
Traversal Time	1200s
Computation Time (2Ghz machine)	3457s
Laser Range Points	4,030,000
Poses	403
Mean Pose Registration	8.6s
Loop Closing Optimization	1.01s
Poses per sub-map when loop closing (n_p)	9

astute reader will notice a discrepancy between the inset of the figure and the plot (where the trajectory undergoes a semicircular perturbation on the western leg of the circuit). This is because the original buildings present in the photograph have now been replaced with a new building — the steps of which can be seen in the inset. Table I gives some pertinent statistics about this system and its application to this data set. The final map contains just over four million range points and four hundred 6DOF vehicle poses. Over 98% of the total run time is spent performing the 3D interpose registration. Overall our current implementation runs at only one third of real-time however we are confident that with further work we can markedly widen the current registration bottleneck. On a 2GHz PC, the most expensive part of the loop closure detection component is the extraction of Harris Affine regions and converting them to SIFT features (around 0.5 seconds per image). For 450 images, vocabulary generation takes a further three minutes and sequence extraction, rank reduction and significance testing takes around 20 seconds.

A. Extension to Multiple Sessions / Vehicles

On a separate occasion another data set was gathered that partially intersected the data set processed to produce the upper loop in Figure 9. The vehicles started from different locations, initialising different global coordinate frames. This is equivalent to having two independent vehicles A and B. In a manner similar to that in [10] we use the vision-based system of Section III to detect shared image sequences. The resulting similarity matrix is shown in Figure 7. Here each element $\mathtt{M}_{i,j}$ is the similarity score between image i from robot A and image j from robot B. Every image from robot A is compared with all images from robot B. Note the off-center dark line indicating overlapping image sequences. In a manner identical to that described in Section III, pairing between image sequences leads (via the estimation of the essential matrix and then scan matching) to an accurate estimate of the transformations between any two vehicle poses lying in areas common to both data sets. Figure 8 shows the matched image sequences from each vehicle. From here it is a simple matter to align both maps/loop in a common coordinate frame. The important point and contribution is that this operation occurs without mutual observation or the existence of a common coordinate frame for all participating robots.

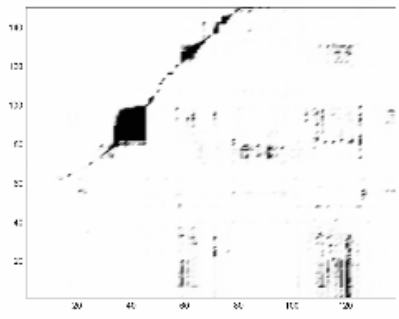

Fig. 7. A visual similarity matrix constructed from the union of two image sequences collected by two robots (before rank reduction). The maximally aligned sequence allows detection of the intersection of the scenes experienced by the independent robots and unification into a common co-ordinate frame.

Fig. 8. Automatically detecting overlapping maps by comparing images sequences observed by different vehicles. The top row is the sequence observed from the vehicle executing the white (upper) loop in Figure 9 while the bottom row is the matching sequence captured by the vehicle that executed the yellow (lower) loop. This matching corresponds to the off diagonal band in Figure 7.

VI. Conclusions and Future Work

This paper has described a SLAM system for outdoor applications and shown it working on a challenging data set in an urban environment. A complementary combination of laser and vision has been used — vision for loop closing and laser data for geometric map building. We showed our system closing a challenging loop — instigated not by geometric considerations but by visual similarity of image sequences. With a potential loop closure detected, constrained non-linear optimisation effected the actual loop closure. To our knowledge this is the first time a 3D ranging sensor has been used in SLAM alongside a vision system for automatic loop closing outdoors. There are however several improvements which we are in the process of researching and implementing. Although convenient, the SLAM formulation we use here is not a efficient one and we would be well served by replacing it with an inverse formulation such as that proposed in [7]. The laser scan matching works well but remains a bottle neck in terms of computation. Finally we are moving towards learning static visual-vocabularies for distinct domains (urban, park-land, indoors etc) and switching between them as the local domain changes. On the grounds that park-like scenes are unlikely to provide evidence for loop closure when working indoors, we propose maintaining a set of domain specific similarity matrices avoiding the cost of maintaining an ever-growing single similarity matrix.

Fig. 9. The resulting estimated map and vehicle trajectory. The inset shows some local detail within the map which contains just over four million range points and the trajectory is decomposed into just over four hundred 6DOF poses. The intersection between the independently executed loops was found in the same way that the loop closures themselves were found - sequential appearance based similarity as described in Section III.

ACKNOWLEDGMENTS

This work is supported in part by the EPRSC through Grant GR/S62215/01 and in part by the Rhodes trust.

REFERENCES

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

[2] J. A. Castellanos, J. M. M. Montiel, J. Neira, and J. D. Tardós, "The SPmap: A probabalistic framework for simultaneous localization and map building," *IEEE Transactions on Robotics and Automation*, vol. 15, no. 5, pp. 948–952, 1999.

[3] J. A. Castellanos, J. Neira, and J. D. Tardós, "Limits to the consistency of ekf-based slam," in *5th IFAC Symp. on Intelligent Autonomous Vehicles (IAV'04)*, Lisbon, Portugal, July 2004.

[4] D. M. Cole and P. M. Newman, "3D SLAM in outdoor environments," *IEEE International Conference on Robotics and Automation*, May 2006.

[5] A. J. Davison and N. Kita, "3D simultaneous localisation and map-building using active vision for a robot moving on undulating terrain," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, Hawaii USA, 11-13 December 2001, pp. 384–391.

[6] R. Eustice, O. Pizarro, and H. Singh, "Visually augmented navigation in an unstructured environment using a delayed state history," in *Proceedings of the 2004 IEEE International Conference on Robotics and Automation*, vol. 1, New Orleans, USA, April 2004, pp. 25–32.

[7] R. Eustice, H. Singh, and J. Leonard, "Exactly sparse delayed-state filters," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA)*, Barcelona Spain, April 2005, pp. 2428–2435.

[8] J. Gutmann and K. Konolige, "Incremental mapping of large cyclic environments," in *Proceedings of the International Symposium on Computational Intelligence in Robotics and Automation (CIRA)*, Monterey California USA, 8-9 November 1999, pp. 318 – 325.

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

[10] K. Ho and P. Newman, "Multiple map intersection detection using visual appearance," in *International Conference on Computational Intelligence, Robotics and Autonomous Systems*, Singapore, December 2005.

[11] K. S. Jones, "Exhaustivity and specificity," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[12] J. Kosecka and X. Yang, "Global localization and relative pose estimation based on scale-invariant features," *Proceedings of the International Conference on Pattern Recognition*, 2004.

[13] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the 7th International Conference on Computer Vision, Kerkyra*, 1999, pp. 1150–1157.

[14] F. Lu and E. Milios, "Globally consistent range scan alignment for environment mapping," *Autonomous Robots*, vol. 4, no. 4, pp. 333–349, 1997.

[15] C. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *International Journal of Computer Vision*, no. 1, pp. 63–86, 2004.

[16] J. Neira and J. D. Tardos, "Data association in stochastic mapping using the joint compatibility test," *IEEE Trans. Robotics and Automation*, vol. 17, no. 6, pp. 890–897, 2001.

[17] P. Newman and K. Ho, "SLAM - Loop Closing with Visually Salient Features," *IEEE International Conference on Robotics and Automation*, 18-22 April 2005.

[18] D. Nister, "An efficient solution to the five-point relative pose problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 26, no. 6, pp. 756–770, June 2004.

[19] E. Royer, L. M., M. Dhome, and T. Chateau, "Towards an alternative GPS sensor in dense urban environment from visual memory," in *Proceedings of British Machine Vision Conference*, London, United Kingdom, 2004.

[20] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proceedings of the International Conference on Computer Vision*, Nice, France, Oct. 2003.

[21] T. Smith and M. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, pp. 195–197, 1981.

[22] H. Surmann, A. Nuchter, and J. Hertzberg, "An autonomous mobile robot with a 3D laser range finder for 3D exploration and digitalization of indoor environments," *Robotics and Autonomous Systems*, vol. 45, pp. 181–198, 2003.

[23] H. Surmann, A. Nuchter, K. Lingemann, and J. Hertzberg, "6D SLAM - preliminary report on closing the loop in six dimensions," in *Proceedings of the 5th IFAC/EURON Symposium on Intelligent Autonomous Vehicles (IAV)*, Lisbon Portugal, 5-7 July 2004.

[24] J. Weingarten and R. Siegwart, "EKF-based 3D SLAM for structured environment reconstruction," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Edmonton Alberta Canada, 2-6 August 2005, pp. 2089 – 2094.