The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Image Augmented Laser Scan Matching for Indoor Dead Reckoning

Nikhil Naikal, John Kua, George Chen, and Avideh Zakhor

Abstract—**Most existing approaches to indoor localization focus on using either cameras or laser scanners as the primary sensor for pose estimation. In scan matching based localization, finding scan point correspondences across scans is challenging as individual scan points lack unique attributes. In camera based localization, one has to deal with images with few or no visual features as well as scale factor ambiguities to recover absolute distances. In this paper, we develop multimodal approaches for two indoor localization problems by fusing a camera and laser scanners in order to alleviate the drawbacks of each individual modality. For our first problem we recover 3 Degrees of Freedom (DoF) of a camera-laser rig on a rolling cart in a 2D plane, by using visual odometry to facilitate scan correspondence estimation. We demonstrate this approach to result in a 0.3% loop closure error for a 60m loop around the interior corridor of a building. In our second problem, we recover 6 DoF of a human operator carrying a backpack system mounted with sensors in 3D, by merging rotation estimates from scan matching and translation estimates from visual odometry, resulting in a 1% loop closure error.**

I. Introduction

Localization in environments with limited global positioning information is a challenging problem. Indoor localization is particularly important in a number of applications such as indoor modeling, and human operator localization in unknown environments. Localization has been primarily studied in the robotics and computer vision communities. In robotics, the focus has been on estimating the joint posterior over the robot's location and the map of the environment using sensors such as wheel encoders, laser scanners and Inertial Measurement Units (IMUs). This is typically referred to as Simultaneous Localization and Mapping (SLAM)[5]. To localize a wheeled robot, simple 2D maps are typically generated using 2D horizontal scanners which serve to both localize the robot and measure depth to obstacles directly. Scan matching based localization approaches such as Iterative Closest Point (ICP) involve computing the most likely alignment between two sets of slightly displaced scans [1]. The open loop nature of the pose integration from ICP and wheel odometry tends to introduce large drifts in the navigation estimates. These estimates can be improved by applying loop closure within a probabilistic framework to estimate the robot's location and the map[4, 6, 7, 8, 9, 10].

The computer vision community has studied pose estimation and Structure from Motion (SfM) for a long time [2, 11, 12, 13, 14]. With a single camera, pose can be estimated only up to an unknown scale factor. This scale is generally determined using GPS waypoints, which makes it inapplicable to indoor environments unless objects of known size are placed in the scene. To resolve this unknown scale factor, stereo camera based approaches have gained popularity, as the extrinsic calibration between the cameras can be used to recover absolute translation parameters [15, 16, 3]. Se *et. al.* present a three camera based stereo system that triangulates SIFT feature correspondences between the cameras to localize a robot mounted with the camera rig [17]. Newman *et. al.* present a system that uses a camera and a 3D laser scanner to localize a vehicle outdoors, with loop closure to enforce global consistency [18].

In this paper, we propose new dead reckoning algorithms that integrate single camera visual odometry, and scan matching to localize a camera and 2D laser scanners. The ultimate goal is to build 3D models of the environment. Even though laser scanners measure the 3D structure of the scene directly and with minimal noise, scan matching is prone to errors in environments with poor geometric features, such as hallways and long corridors. Camera images, on the other hand, capture color and texture from which visual correspondences can be found across images. Visual odometry techniques perform poorly when there are few, or no visual features in the images. In this paper, we show that fusing camera and laser scanners is likely to overcome some of the above shortcomings of each in order to improve localization accuracy. Specifically, we consider two indoor localization problems. The first one deals with recovering the 3 DoF motion parameters of a sensor rig mounted on a rolling cart in the 2D plane; the second one deals with full 6 DoF localization of a backpack of sensors being carried by a human operator in 3D.

For the 2D case on a cart, we mount a horizontal and a vertical laser scanner on the rig and strap down a side looking camera. The rig is placed on a wheeled cart, and is pushed around the interior corridors of a building. The purpose of the horizontal laser scanner is to localize, while the vertical scanner provides a dense point cloud of the environment for geometry modeling purposes. The camera serves a dual purpose in that it is used for both localization and providing color and texture to the 3D models. We introduce a Visual Odometry aided Scan Matching (VOSM) method that uses visual odometry to determine the camera pose between successive images, which in turn aids in determining scan correspondence estimates across images.

For the 3D localization problem, we mount the rig on a backpack that is carried by a human operator. Specifically, we have mounted three 2D laser scanners orthogonal to each other in order to estimate the yaw, pitch, and roll. We then use these angle estimates within a visual odometry algorithm

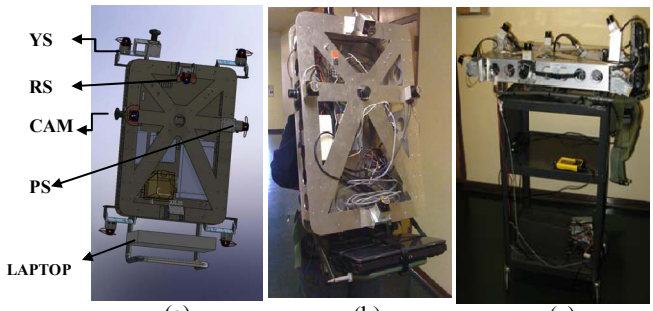
Manuscript received March 1, 2009. This work was supported by the ARO grant W911NF-07-1-0471 and HSN MURI grant W911NF-06-1-0076.

Nikhil Naikal, John Kua, and Goerge Chen are with the Electrical Engineering Department, University of California, Berkeley, USA(email: {nnaikal, jkua, gchen}@eecs.berkeley.edu).

Avideh Zakhor is a Professor at the Electrical Engineering Department, University of California, Berkeley, USA(email: avz@eecs.berkeley.edu).

to compute all the 6 pose transformation parameters in 3D. Fig. 1(a) depicts the conceptual CAD model of the backpack system, Fig. 1(b) shows the backpack system being carried by a human operator, and Fig. 1(c) shows the system placed on a rolling cart.

This paper is organized as follows. In Section II we present our extrinsic calibration method to find the relative orientation between a 2D horizontal laser scanner and a camera. In Section III we provide an overview of existing pose estimation methods for standard visual odometry with specific implementation details. In Section IV we describe our VOSM algorithm for 2D dead-reckoning and characterize its performance on an indoor dataset. In section V we introduce a combined laser-camera dead reckoning algorithm in 3D, and characterize its performance against ground truth collected in an indoor corridor with minimal clutter and obstacles. Conclusions and future research are presented in Section VI.



(a) (b) (c)

Fig 1- The data acquisition system; (a) CAD model of backpack system; YS, PS, and RS are the short range laser scanners for estimating yaw, pitch, and roll, and CAM is the camera; (b) the assembled backpack system carried by a human operator; (c) the system placed on a cart.

II. Extrinsic Sensor Calibration

The relative rigid transformation between the camera and the laser scanner is needed to effectively fuse the two sensors. We determine the camera's internal parameters using the Caltech camera calibration toolbox [19]. We compute the extrinsic calibration between the camera-laser pair only once, as the sensors are rigidly mounted relative to each other. Using the pinhole camera model, a 3D point in camera coordinates, $\boldsymbol{p}^c = [x^c, y^c, z^c]^T$, is represented in image coordinates as,

$$\boldsymbol{p} = [p_x \quad p_y \quad 1]^T = \boldsymbol{K}\,[x^c/\mathrm{z}^c \quad y^c/z^c \quad 1]^T \qquad (1)$$

where $\boldsymbol{K}$ is the intrinsic camera calibration matrix, and $\boldsymbol{p}$ is the image pixel location of point $\boldsymbol{p}^c$. Thus, the unit vector of the directional line from the camera center to $\boldsymbol{p}^c$ is,

$$\hat{\boldsymbol{p}}^c = \boldsymbol{K}^{-1}\boldsymbol{p}/\|\boldsymbol{K}^{-1}\boldsymbol{p}\| \qquad (2)$$

The laser scanner measures a 2D slice of the scene; thus, in laser coordinates a scan point is assumed to lie on the plane $Z = 0$, and is represented by $\boldsymbol{p}^l = [x^l, y^l, 0]^T$. We begin by manually choosing three (laser point, image vector) pairs, i.e., $([\boldsymbol{p}_1^l, \boldsymbol{p}_2^l, \boldsymbol{p}_3^l] \leftrightarrow [\hat{\boldsymbol{p}}_1^c, \hat{\boldsymbol{p}}_2^c, \hat{\boldsymbol{p}}_3^c])$, corresponding to three world points, $[\boldsymbol{P_1}, \boldsymbol{P_2}, \boldsymbol{P_3}]$. These pairs are used by the 3-point algorithm (p3p) to determine the distance to the world points from the camera center, thus recovering their position in camera coordinates [20]. The relative pose between the sensors is now obtained by applying Horn's method to the three point pairs in laser and recovered camera coordinates [21]. The p3p algorithm requires the distance between the 3D world points to be known. This is obtained by computing the Euclidean distance between pairs of laser points.

We use a thin rectangular box placed at the height of the laser as the calibration target. Laser scan point-camera pixel correspondences are obtained by manually selecting the two ends of the box. Calibration sets consisting of scans and images of the target are collected from 20 to 30 different locations by moving the sensor platform. We have implemented our extrinsic calibration process within a RANSAC framework, where we choose three point-pixel pairs at a time to generate a pose hypothesis, and evaluate it on all the calibration sets. The winning hypothesis is further refined using iterative nonlinear methods, where the objective function being minimized is the error between scan-point back-projections and true pixel locations of the points in all the calibration images. To project laser scans onto images, we first transform each scan point $\boldsymbol{p}^l$ to the camera coordinate frame using the estimated rotation and translation from laser to camera frame of reference, i.e. $[\boldsymbol{R}_l^c, \boldsymbol{t}_l^c]$. We then find the image coordinates of the point using Eqn. 1. Fig. 2 shows a scatter plot representing the error between laser point back-projections and true pixel locations of the corresponding image points for all 30 calibration sets. On average, there is a 6 pixel error in laser scan back-projection on to camera images.

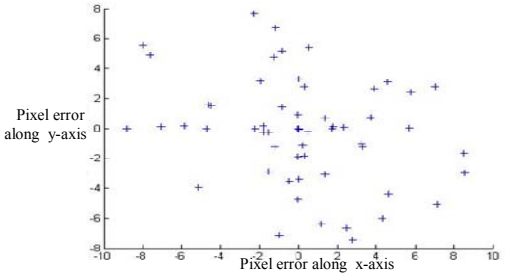


Fig. 2. Scatter plot of error between manually chosen image pixel locations and corresponding scan point back-projections onto images found using computed laser-camera extrinsic parameters.

III. Visual Odometry

Image sequences from a camera could potentially provide sufficient information to determine a camera's trajectory. In the visual odometry setting, features in images are tracked between frames to determine the pose of an internally calibrated camera from the visual feature correspondences. The epipolar constraint between two overlapping camera views are enforced by the essential matrix, $\boldsymbol{E}$, such that, for any two calibrated point correspondences $\boldsymbol{p} \leftrightarrow \boldsymbol{p}'$, we have,

$$(\boldsymbol{K}^{-1}\boldsymbol{p}')^T\boldsymbol{E}(\boldsymbol{K}^{-1}\boldsymbol{p}) = 0 \qquad (3)$$

The 5-Point algorithm can be used to determine the essential matrix in scenes with planar degeneracies which are ubiquitous in indoor environments [13]. As the name suggests the algorithm determines $\boldsymbol{E}$ given 5 image feature

correspondences. The epipolar geometry computation is in general most accurate when sufficient motion occurs between two image frames. Hence, we choose to detect and track SIFT features across successive images until the number of correspondences falls below a preset threshold [22]. We then compute the essential matrix between the first and last image in the tracked image sequence with the five-point algorithm within a preemptive RANSAC routine. Finally, we apply nonlinear iterative refinement to find the best pose transformation that minimizes the Sampson reprojection error of the features between the two images. The convenient structure of the essential matrix $\boldsymbol{E}$ allows it to be decomposed into a rotation and translation because,

$$\boldsymbol{E} = [\hat{\boldsymbol{t}}_c]_\times \boldsymbol{R}_c, \tag{4}$$

where $[\boldsymbol{R}_c, \hat{\boldsymbol{t}}_c]$ represent the camera rotation and unit translation direction, and $[.]_\times$ is the cross product.

IV. Visual Odometry aided Scan Matching (VOSM)

In static environments with sufficient geometric features, such as walls at different angles and other obstacles, point-wise scan matching can be used to determine the ego-motion of the moving horizontal laser scanner. ICP [1] is the most popular scan matching algorithm which iteratively computes the scan transformation, $[\boldsymbol{R}_l, \boldsymbol{t}_l]$, by minimizing the squared distance between each of the N points in the first scan, $\boldsymbol{m}$, and their nearest neighbors in the second scan, $\boldsymbol{d}$, i.e.,

$$\min_{\boldsymbol{R}_l, \boldsymbol{t}_l} \sum_{\mathrm{N}} \|\boldsymbol{m}_i - \boldsymbol{R}_l \boldsymbol{d}_i - \boldsymbol{t}_l\|^2 . \tag{5}$$

A na ve nearest neighbor approach to find point correspondences is likely to fail when the environment being scanned has few geometric features. The basic idea behind VOSM is to take advantage of feature rich visual imagery to compensate for the lack of geometric features in scenes. Specifically, our approach in VOSM is to assign scan point correspondences across successive scans by using the rotation and translation from visual odometry. We use these correspondences to compute the transformation between the two successive scans within a RANSAC framework. The details of the VOSM algorithm are provided in the remainder of this section.

A. Image Based Nearest Neighbor Search

We project laser scans of a scene from two different locations onto their corresponding images. The scan projection tracker finds the best scan point correspondences across the two images as follows.

1. Two successive laser scans, $[\boldsymbol{L}_t, \boldsymbol{L}_{t+1}]$, are projected onto their corresponding images, $[\boldsymbol{I}_t, \boldsymbol{I}_{t+1}]$.
2. The rotation and translation estimates from visual odometry are used to warp image $\boldsymbol{I}_{t+1}$ into image $\boldsymbol{I}_t$'s view, and to determine a search window across the two images to find scan correspondences.
3. Image patches are extracted around each scan point projection in images $\boldsymbol{I}_t$ and $\boldsymbol{I}_{t+1}$ in order to find patch correspondences across images by minimizing the bi-directional Sum of Absolute Difference (SAD) within the search window established by visual odometry.

In our experiments, the optimal window patch size was empirically set to 31x31 pixels, and the search window was restricted to be 50x50 pixels.

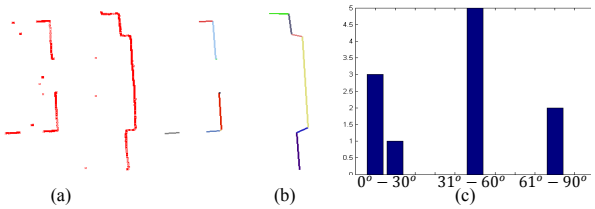
B. Robust Scan Matching

Once the scan point correspondences are found using images, the rigid transformation between the two sets of scan points can be obtained directly without any iterative scheme. However, to improve the robustness of the matching process, we adopt a RANSAC based approach in which two sets of candidate point matches are randomly selected, and a pose hypothesis is computed. This candidate hypothesis is evaluated on all the scan point correspondences, and a score is assigned to it. The hypothesis evaluation scheme is determined based on the angular distribution of lines in the scan. At the end of the routine, the winning hypothesis is chosen as the one with the highest score. To determine the hypothesis evaluation metric, lines are extracted in each scan, and an angle histogram is computed, with 10° bins as shown in Fig. 3(c). Each line's angle relative to the scanner is determined from its slope. If the angular distribution of scanned lines is sufficiently wide, then a laser based metric to evaluate the RANSAC hypothesis is instantiated. In this case, each candidate pose hypothesis is scored inversely to the alignment error between the second scan and the first scan transformed with the hypothesis. Fig. 3(a) shows a typical scene where the laser based evaluation metric is used since there is a wide distribution of lines across many angles as seen in Fig. 3(b).

On the other hand, if the angular distribution of scan lines is narrow, then an image based evaluation method is used. Specifically, for each subset of two point correspondences, a pose hypothesis is generated. With this hypothesis, the first scan, $\boldsymbol{L}_t$, is transformed and projected onto the warped second image $\boldsymbol{I}_{t+1}$. The SAD of image patches around each projected scan point of $\boldsymbol{L}_t$ between the first and second image, i.e. $\boldsymbol{I}_t$ and $\boldsymbol{I}_{t+1}$, is computed. The hypothesis score assigned is inversely proportional to the mean of the SAD error of all image patches. A final stage of ICP is performed to refine the computed pose transformations.

C. Scale Computation

The true scale in the translation, $\hat{\boldsymbol{t}}_c$, computed via visual odometry is determined as follows. For the first pair of images in the tracked image sequence, the search window for the nearest neighbor algorithm presented in section IV-A is set as the size of the entire image. The 3D coordinates of a single point, $\boldsymbol{P}$, is obtained from the laser scanner, and its location in the first and last image in the tracked image sequence are obtained from the image patch correspondence algorithm. This image correspondence pair is triangulated with the current camera pose estimate, $[\boldsymbol{R}_c, \hat{\boldsymbol{t}}_c]$, to determine the scaled coordinates of the point, i.e., $\widehat{\boldsymbol{P}}$. The scale in the translation is then obtained directly as,



(a) (b) (c)

Fig. 3. (a) Sample laser scan; (b) lines extracted from the scan in (a); (c) angle distribution of the lines in (b).

$$s = \|\boldsymbol{P}\| / \|\hat{\boldsymbol{P}}\|, \tag{6}$$

where, $\|.\|$, is the Euclidean norm. The triangulation procedure is described in detail in [13]. After this form of initialization, the scale in translation for subsequent image pairs, $[\boldsymbol{I}_t, \boldsymbol{I}_{t+1}]$, is obtained in a boot-strapped manner, by putting the triangulated features, $\hat{\boldsymbol{P}}_t$, in the coordinate system of the triangulated features, $\boldsymbol{P}_{t-1}$, in the previous image pair, $[\boldsymbol{I}_{t-1}, \boldsymbol{I}_t]$. With this computed scale, the corresponding location in the second view, $\boldsymbol{P}_{t+1}$, of a point $\boldsymbol{P}_t$ in the first view is found using,

$$\boldsymbol{P}_{t+1} = \boldsymbol{R}_c \boldsymbol{P}_t + s_t \hat{\boldsymbol{t}}_c \tag{7}$$

where, $[\boldsymbol{R}_c, \hat{\boldsymbol{t}}_c]$, are the camera rotation and unit translation obtained from visual odometry, and s_t is the translation scale at the current time, t. The search window for the image patch matching algorithm is centered around the pixel back-projection of $\boldsymbol{P}_{t+1}$ onto the warped image $\boldsymbol{I}_{t+1}$.

D. Algorithm

Fig. 4 shows the flowchart of the VOSM algorithm. Since the laser scanner and camera operate at different frame rates, the two sensors are initially synchronized. The laser scans are then transformed to camera coordinates with the extrinsic calibration computed earlier. Two successive images and their corresponding laser scans are input into the visual odometry and scan matching sub-systems. The visual odometry system computes the camera rotation and translation, with the scale in translation computed using the p3p bootstrapping method. This camera pose matrix is used to transform the laser points from the first image's coordinate system to the second image's coordinates, and to warp the second image to be consistent with the first camera's view. Projecting these transformed laser points onto the warped image plane in the second view provides a local search region to find scan point correspondences in the images. The patch based search method described in Section IV-A is employed to find the best matches by minimizing the SAD of image patches around scan point projections in the two images but searching only within the local search window. Once correspondences are found, the robust RANSAC based method described in Section IV-B determines the pose transformation. The bootstrapped method of computing scale, as explained in section IV-C, tends to accumulate errors over time. Thus, scale in the current translation is re-computed using Eqn. 6, and the SIFT features are re-triangulated with the pose estimates from visual odometry.

E. VOSM Experimental Results

We compare the accuracy of the VOSM algorithm presented in this paper with the ground truth collected by an Applanix position and orientation system used for land surveying. This is an aided inertial navigation system consisting of a navigation computer and a strap down navigation-grade Honeywell HG9900 IMU. The HG9900 combines three ring laser gyros with bias stability of less than 0.003deg/hr, and three precision accelerometers with bias of less than 0.245 mm/s^2. For our indoor experiments, we utilized a pre-surveyed control point as a global position reference. Navigation precision is improved by the use of zero-velocity updates (ZUPTs), which allow for accumulated biases in the IMU to be estimated, and any velocity drift to be corrected. These ZUPTs manifest as discontinuities in the ground truth paths of Fig. 5 to be discussed shortly.

Ground truth comparison of VOSM for a 60m loop inside a corridor of a building on the UC Berkeley campus is shown in Fig. 5(a). The raw visual odometry and ICP results are plotted against ground truth in Fig. 5(b). As seen, the VOSM reconstructed path is in close agreement with the ground truth, while VO and ICP visibly deviate from the ground truth. Figures 5(c)-5(e) compare the rotation and translations from VOSM against the ground truth. As seen, there is close agreement between the ground truth and VOSM values. The loop closure error for all schemes is shown in Table-1. As expected, the loop closure error is the lowest for VOSM at 18cm, or 0.3% of the traversed path. In contrast, Oskiper *et. al.* [3] have reported on a more elaborate system consisting of two stereo camera pairs and an IMU to obtain between 0.5% to 1% loop closure error. In addition to loop closure error, we have also computed the average position error for the various algorithms by determining the distance between the ground truth position and the position computed by each algorithm at each time step in the 3rd column of Table-1. As expected, VOSM has an average position error that is 15 times smaller than that of ICP and 7 times smaller than that of VO. The 3rd and 4th columns of Table-1 confirm the plots in Fig. 5 showing that yaw and translation parameters for VOSM have lower RMS error compared to VO and ICP.

Dead Reckoning Method	**Loop Closure Error (m)**	**Average Position Error (m)**	**RMS Yaw Error (degrees)**	**RMS Translation Error (m)**
ICP	0.24	2.36	0.17	X: 0.05 Y: 0.03
VO	1.23	1.09	0.38	X: 0.04 Y: 0.06
VOSM	**0.18**	**0.15**	**0.11**	**X: 0.02** **Y:0.02**

Table-1: A comparison of the mean position and loop closure errors for ICP, Visual Odometry (VO), and VOSM.

V. BACKPACK SYSTEM FOR INDOOR MODELING

VOSM is essentially designed to provide navigation estimates for a wheeled system that has 3 degrees of freedom. The algorithm, however, does not apply to a system that has a non-zero pitch and roll, such as a backpack system carried by a human operator. Visual odometry provides pose estimates of a traversing camera in 3D and

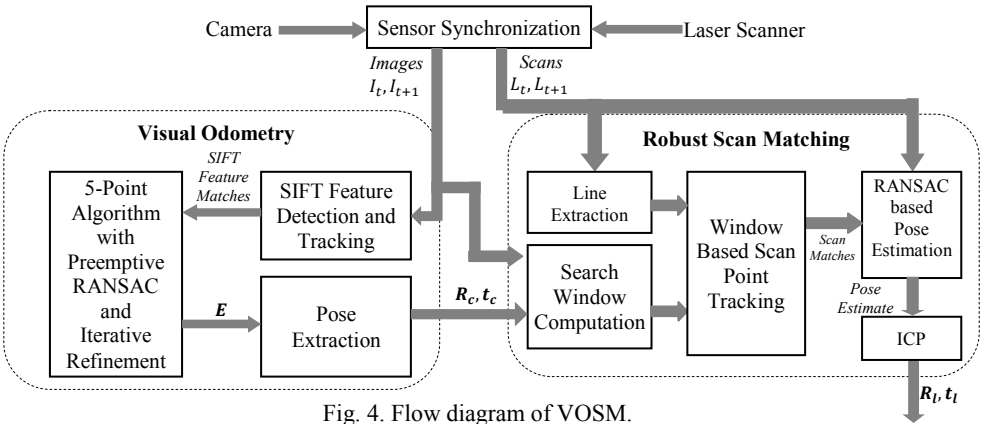


Fig. 4. Flow diagram of VOSM.

performs well when the camera has a smooth trajectory. However, the change in dynamics of the human gait while walking is quite complex and as such, we have empirically found that visual odometry, by itself does not perform well. Further, with a side facing camera, features are only tracked over a short sequence of images, and no long term feature tracks are available to accurately determine the three rotation parameters, thus leading to large accumulation of error over time.

A horizontally mounted laser scanner, on the other hand, measures the absolute depth to objects in the scene. Further, the rotation computed by scan matching is generally more accurate than visual odometry in indoor scenes, since the sensor's scanning rate is much faster than the speed at which a human operator traverses the environment. Thus, we have opted to mount 3 laser scanners orthogonal to each other in our backpack system in order to measure the rotation in the 3 independent axes namely X, Y, and Z, as shown in Fig. 1(a). With these initial rotation estimates, the translation vector is recovered using camera images as explained shortly.

A. ICP Aided Visual Odometry (ICP-VO)

We transform ICP rotation estimates from the 3 orthogonal laser scanners to camera coordinates to construct a full 3D rotation matrix to represent the rotation between a pair of successive images, $\boldsymbol{I_t}$ and $\boldsymbol{I_{t+1}}$.. The SIFT feature correspondences between this pair of images is obtained from the feature tracker. We let,

$$\boldsymbol{f_i'} = \boldsymbol{K^{-1}p_i'} \text{ and } \boldsymbol{f_i} = \boldsymbol{K^{-1}p_i} \tag{8}$$

represent the vectors along which a feature correspondence pair lies in the two images. Substituting Eqns. 4 and 8 in Eqn. 3 we obtain,

$$\boldsymbol{f_i'}^T.[\hat{\boldsymbol{t}}_c]_\times \boldsymbol{R}_c.\boldsymbol{f_i} = 0, \tag{9}$$

By defining $\boldsymbol{f_i'} = [f_x', f_y', 1]^t$, and reordering terms we have,

$$\begin{bmatrix} -f_y'.\boldsymbol{r_3}.\boldsymbol{f_i} + \boldsymbol{r_2}.\boldsymbol{f_i} \\ f_x'.\boldsymbol{r_3}.\boldsymbol{f_i} - \boldsymbol{r_1}.\boldsymbol{f_i} \\ f_y'.\boldsymbol{r_1}.\boldsymbol{f_i} - f_x'.\boldsymbol{r_2}.\boldsymbol{f_i} \end{bmatrix}^T . \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} = 0, \tag{10}$$

where, $\boldsymbol{r_1}, \boldsymbol{r_2}, \boldsymbol{r_3}$ are the 3 rows of the rotation matrix, $\boldsymbol{R_c}$, that is composed from the 3 Euler angles obtained by performing scan matching on the 3 orthogonal laser scanners. With all the available feature correspondences between the two images, the null space of the matrix in Eqn. 10 is spanned by the translation vector, $\hat{\boldsymbol{t}}_c = [t_x, t_y, t_z]^T$.

It is important to note that Eqn. 10 has only 2 degrees of freedom, as the translation can be computed only up to an unknown scale factor. Thus, only 2 feature point correspondences are sufficient to find the translation. However, feature correspondences from the SIFT feature tracker could have outliers that degrade the translation solution. Thus, we have implemented a 2-point RANSAC procedure that computes a translation hypothesis using 2 randomly chosen point correspondences, and scores the hypothesis based on the reprojection error of all feature correspondences between the two images. The winning pose estimates are then further refined iteratively to find the best solution that minimizes the Sampson reprojection error of the feature correspondences between the 2 images. With our current backpack configuration, multiple laser scanners sweep the environment as the human operator traverses it, and absolute depth can be assigned to SIFT features in the images when laser scan points project to within a few pixels of the feature location in an image. The true depth of the SIFT feature is used to directly compute the scale in translation using Eqn. 6.

B. Results

To evaluate the performance of the ICP-VO algorithm, two data sets were collected in the interior corridor of the Electrical Engineering building on UC Berkeley campus. The results of the algorithm on the first dataset are compared against ground truth from the HG9900, in Fig. 6(a). The initial 6 DoF pose from the IMU is applied to the ICP-VO reconstructed path in order to compare the paths. Unlike VOSM, the recovered path by ICP-VO is truly in 3D since it recovers 6DoF pose rather than 3DoF. As seen, the ICP-VO path is in close agreement with the ground truth in the x-y plane, but not along the z-axis; this can be attributed to errors in roll and pitch. We have found that by applying local bundle adjustment techniques such as the one described in [14], we can significantly reduce the error in pitch and roll, Y, and Z, thereby improving the accuracy of the reconstructed path along the z-axis. For comparison, Fig. 6(b) shows the reconstructed path by VO and ICP against ground truth. As seen, VO results in large error in the z-axis and ICP has large error along the x-axis. Figs. 6(c) through 6(h) compare the six pose parameters against the ground truth from the Honeywell IMU. As seen, there is close agreement between the pose transformation values and the

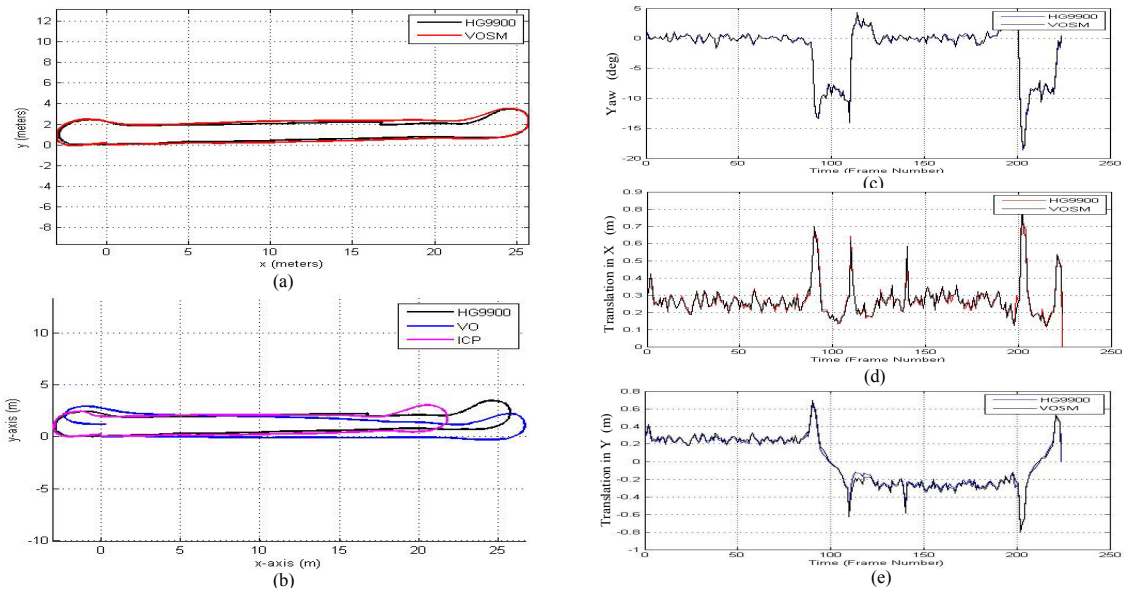


Fig. 5. (a) Reconstructed VOSM path, in red, and ground truth in black; (b) the raw ICP path, in pink, and visual odometry path, in blue against ground truth, in black; (c) comparison of VOSM yaw in black, against ground truth in blue; (d) comparison of VOSM x-translation in black, against ground truth in red; (e) comparison of VOSM y-translation in black, against ground truth in blue.

ground truth. For the second data set, a different operator carried the backpack, with different walking dynamics, and larger incremental rotations occurring at faster time scales. This resulted in slightly larger loop closure and path errors as seen in Fig 7(a).

The average path and loop closure errors for the two data sets are reported in Table-2. As seen the loop closure, and average position errors for ICP-VO is smaller than VO and ICP for both sets. The 4th and 5th columns of Table 2 show the RMS and peak errors for all pose parameters. The RMS error for yaw is considerably smaller for ICP-VO than for VO and ICP, and the peak error for all 6 pose parameters is considerably smaller for ICP-VO than for VO and ICP.

The loop closure error for the VOSM algorithm which only recovers x, y and yaw in the 2D plane is 1.29m (1.23 m) for set-1 (set-2). Similarly, the average position errors are 0.52m (0.72m) for set-1 (set-2). Thus, for backpack data, a truly 3D algorithm such as ICP-VO not only recovers all 6 degrees of freedom, but also results in smaller loop closure error.

VI. Conclusions and Future Work

In this paper two image augmented laser scan matching algorithms have been presented for indoor dead-reckoning. The VOSM algorithm is predominantly a 2D algorithm that efficiently uses images to recover 3DoF poses in a 2D plane. It has been tested in the interior corridor of a building, and results in a 0.3% loop closure error. This is better than the loop closure error obtained in [3] for a combined indoor-outdoor path. The ICP-VO algorithm presented also fuses laser scan matching with image based pose estimated in a 3D framework, and provides an accurate method of dead-reckoning in 3D, with loop closure errors of about 1% of the traversed path. Future work involves loop closure detection, and enforcing global consistency using a graph-SLAM framework. Ultimately, we plan on applying our proposed algorithms to localize our backpack mounted with laser scanners and cameras for 3D indoor modeling.

	Loop Closure Error(m)	Average Position Error(m)	Rotation Error (Y/P/R) (deg)	Translation Error (X/Y/Z)(m)
Set 1				
VO	3.99	2.88	*RMS Error* 1.81/0.61/0.94	*RMS Error* 0.07/0.02/0.05
			Peak Error 4.9/3.4/3.92	*Peak Error* 0.83/0.33/0.37
ICP	1.97	3.53	*RMS Error* 0.47/0.67/0.81	*RMS Error* 0.1/0.04/0.05
			Peak Error 3.71/3.39/6.86	*Peak Error* 0.54/0.26/0.2
ICP-VO	0.65	0.57	*RMS Error* 0.30/1.65/1.35	*RMS Error* 0.05/0.04/0.05
			Peak Error 2.5/1.92/2	*Peak Error* 0.24/0.16/0.17
Set 2				
VO	3.61	1.09	*RMS Error* 1.67/0.89/0.56	*RMS Error* 0.13/0.04/0.11
			Peak Error 7.1/3.15/7.16	*Peak Error* 1.02/0.96/0.31
ICP	4.88	8.05	*RMS Error* 0.79/0.95/0.94	*RMS Error* 0.17/0.06/0.09
			Peak Error 5.28/6.62/6.2	*Peak Error* 1.7/0.94/0.41
ICP-VO	0.69	0.86	*RMS Error* 0.77/1.19/0.93	*RMS Error* 0.08/0.05/0.10
			Peak Error 3.5/2.33/2.75	*Peak Error* 0.29/0.48/0.25

Table 2 – Performance of VO and ICP-VO on indoor data sets.

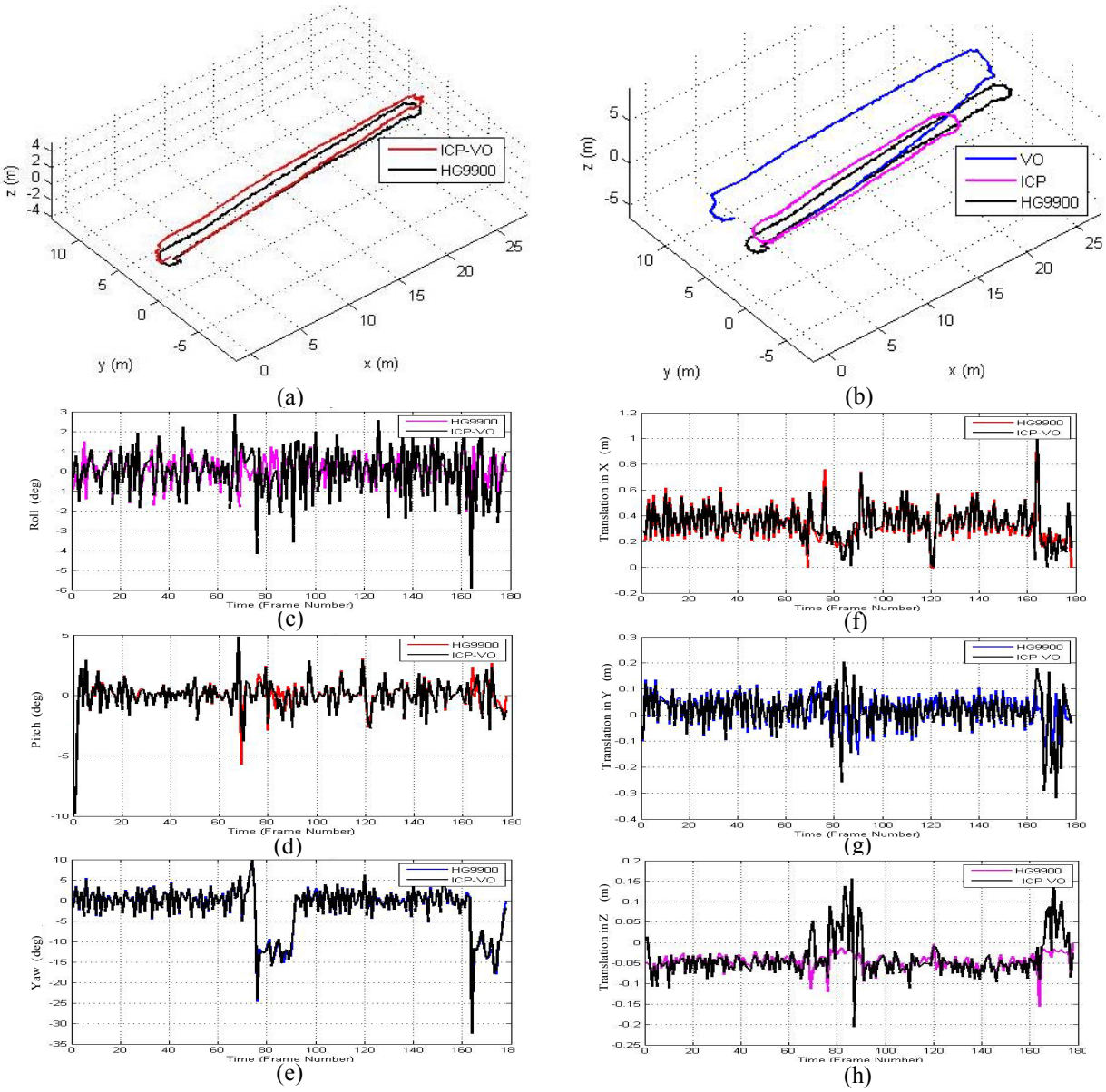


Fig. 6 (a). Reconstructed path of the ICP-VO algorithm in red, against ground truth in black for Set 1; (b) reconstructed VO path in blue, and ICP in pink, against ground truth in black; (c)-(e) computed roll, pitch and yaw from ICP-VO against ground truth; (f)-(h) computed delta translations in the x, y and z directions against ground truth.

VII. REFERENCES

[1] F. Lu and E. Milios, Robot *pose estimation in unknown environments by matching 2D range scans*, Jnl. of Intelligent and Robotic Systems, 1997.

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

[3] T. Oskiper, Z. Zhu, S. Samarasekara and R. Kumar, *Visual Odometry System Using Multiple Stereo Cameras and Inertial Measurement Unit*, CVPR, 2007.

[4] A. Mourikis and S. Roumeliotis, *A Multi-State Constraint Kalman Filter for Vision-aided Inertial Navigation*, ICRA 2007. [5] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*, MIT Press, Cambridge, MA, 2005.

[5] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*, MIT Press, Cambridge, MA, 2005.

[6] H. Durrant-Whyte and T. Bailey, *Simultaneous Localization and Mapping (SLAM): Part I The Essential Algorithms,*. Robotics and Automation Magazine, 2006.

[7] S. Thrun, W. Burgard, and D. Fox, *A real-time algorithm for mobile robot mapping with applications to multirobot and 3d mapping*. Proc. ICRA , 2000.

[8] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, *Fast-SLAM 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges,* International Joint Conference on Artificial Intelligence, 2003.

[9] A. Davison, *Real-time simultaneous localisation and mapping with a single camera*, ICCV, 2003.

[10] E. Eade and T. Drummond, *Scalable Monocular SLAM*, Proc. CVPR, 2006.

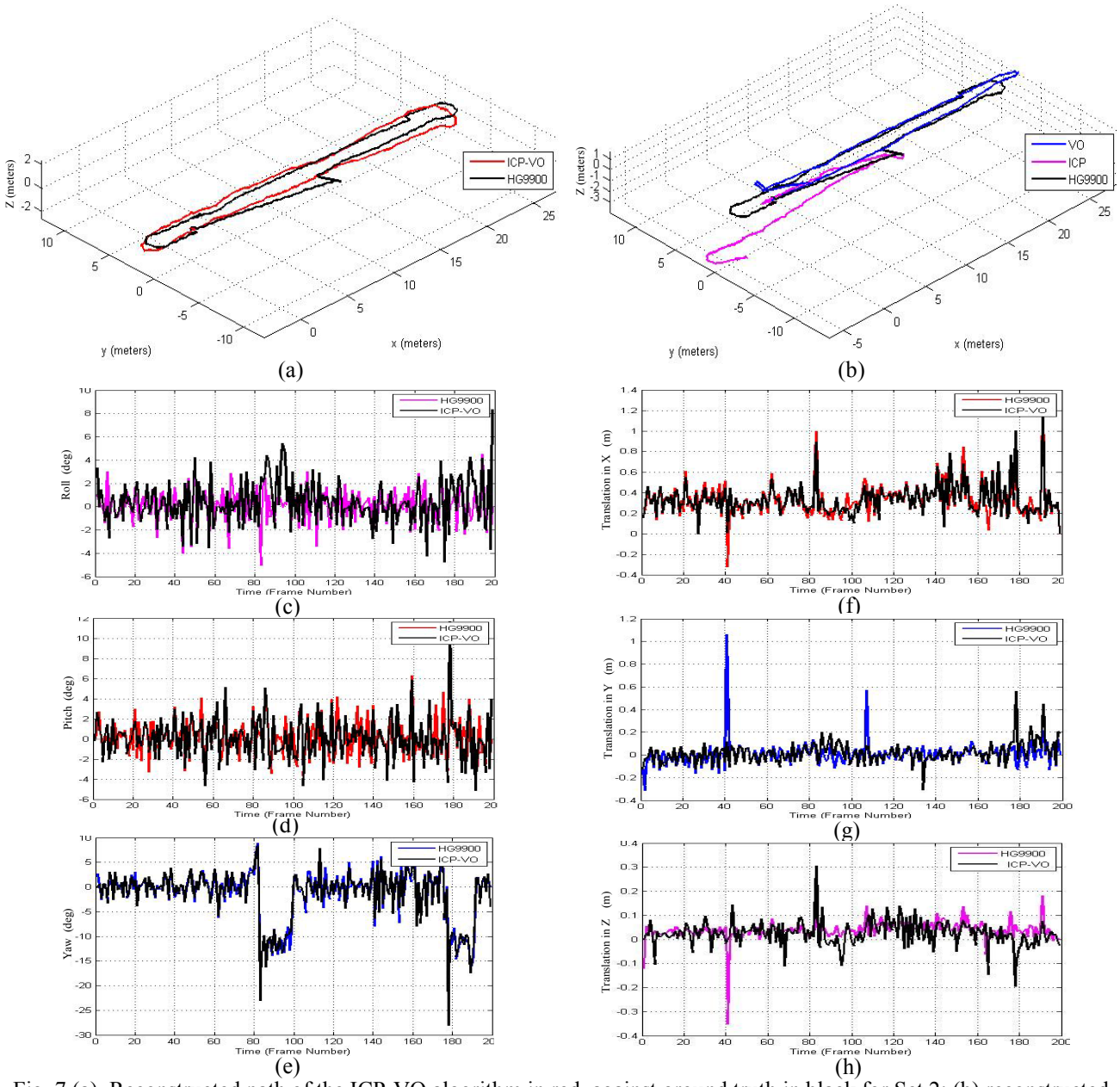


Fig. 7 (a). Reconstructed path of the ICP-VO algorithm in red, against ground truth in black for Set 2; (b) reconstructed VO path in blue, and ICP path in pink, against ground truth in black; (c)-(e) computed roll, pitch and yaw from ICP-VO against ground truth; (f)-(h) computed delta translations in the x, y and z directions against ground truth.

[11] D. Nister *Automatic Dense Reconstruction from Uncalibrated Video Sequences*, PhD Thesis, University of Stockholm, 2001

[12] O.D. Faugeras and Q.T. Luong, *The Geometry of Multiple Images,* The MIT Press, 2001.

[13] D. Nister, *An efficient solution to the five-point relative pose problem* IEEE Transactions on PAMI, 2004.

[14] E. Mouragnon , M. Lhuillier , M. Dhome , F. Dekeyser, and P. Sayd, *Real Time Localization and 3D Reconstruction*, Proc. CVPR, 2006.

[15] D. Nister, O. Naroditsky, and J. Bergen, *Visual odometry*, in Proc. CVPR, 2004.

[16] M. Agrawal and K. Konolige, *Real-time localization in outdoor environments using stereo vision and inexpensive GPS*, Proc. ICPR, 2006.

[17] S. Se, D. Lowe, and J. Little, *Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks*, International Journal of Robotics, 2002.

[18] P. Newman, D. Cole, and K. Ho. *Outdoor SLAM using Visual Appearance and Laser Ranging*. Proc. ICRA, 2006.

[19] http://www.vision.caltech.edu/bouguetj/calib_doc/

[20] R.M. Haralick, C.N. Lee, K. Ottenberg, and M. Nolle, *Review and analysis of solutions of the three point perspective pose estimation problem*, IJCV 1994.

[21] B.K.P. Horn, *Closed Form Solution of Absolute Orientation Using Unit Quaternions*, Jnl of Opt. Soc. of America, 1987.

[22] D. Lowe, *Distinctive Image Features from Scale-Invariant keypoints,* IJCV, 2004.