Image Augmented Laser Scan Matching for Indoor Localization

Nikhil Naikal
Avideh Zakhor
John Kua

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2009-35
http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-35.html

March 2, 2009

Image Augmented Laser Scan Matching for Indoor Localization

Nikhil Naikal, John Kua and Avideh Zakhor
Video and Image Processing Lab
University of California, Berkeley
{nnaikal, jkua, avz}@eecs.berkeley.edu

Abstract

Indoor localization is a challenging problem addressed extensively by both robotics and computer vision communities. Most existing approaches focus on using either cameras or laser scanners as the primary sensor for pose estimation.In laser scan matching based localization, finding scan point correspondences across scans is a challenging problem as individual scan points lack unique attributes. In camera based localization, one has to deal with images with little or no visual features as well as scale factor ambiguities to recover absolute distances. In this paper, we develop a multimodal approach for indoor localization by fusing a camera and a laser scanner in order to alleviate the drawbacks of each individual modality. Specifically, we use structure from motion to estimate the pose of a moving camera-laser rig which is subsequently used to compute piecewise homographies for planes in the scene scanned by the laser scanner. The homographies provide scan correspondence estimates which are refined using a window based search method for scan point projections on the images. We have demonstrated our proposed system, consisting of a laser scanner and a camera, to result in a 0.3% loop closure error for a 60m loop around the interior corridor of a building.

1. Introduction

Localization in environments with limited global positioning information remains a challenging problem. Indoor localization is a particularly important problem with a number of applications such as indoor modeling, and human operator localization in unknown environments. Localization has been primarily studied by the robotics and computer vision research 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 problem 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 laser scanners which serve to both localize the robot and measure depth to obstacles directly. Laser scan matching based localization approaches involve computing the most likely alignment between two sets of slightly displaced laser scans. The Iterative Closest Point (ICP) algorithm [1] is the most extensively used scan correlation algorithm in robotics. 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 using a dynamic motion model for the robot, and by applying probabilistic methods to estimate the robot's location and the map. Thrun *et. al.*[7] use an expectation maximization approach to solve the SLAM problem on robots with wheel encoders and laser scanners. Other approaches to SLAM tend to rely heavily on Extended Kalman Filters (EKFs)[6], since the process of tracking the robot's pose and position of the geometric features in the scene can be elegantly represented within the EKF estimation framework. However, this method becomes computationally intensive while traversing large distances, and as such, different speedups have been proposed [8].

Cameras are used less often than range scanners in solving the SLAM problem, due to associated computational complexity and real time requirements. However, with increasing processor speeds, vision based SLAM has become more feasible over the past decade. Davison[9] has proposed a real time vision based EKF-SLAM algorithm to localize a monocular camera moving with a smooth trajectory. The computational complexity of the EKF renders his algorithm useful only for mapping small environments. Eade and Drummond present a particle filter based version of monocular SLAM which can scale to large environments[10]. Mourikis and Roumiliotis speed up the vision based EKF-SLAM algorithm by developing a measurement model that does not include the feature positions in the filter state vector[4]. These vision based SLAM methods rely on observing the same visual features over a large set of images. This is generally not possible with a side looking camera traversing indoor environments, as features become unavailable after few frames.

The computer vision community has studied the localization problem in the context of Structure from Motion (SfM) [2, 11, 12, 13]. The geometric relationship between the images of a scene viewed from two different locations provides the necessary constraints to determine the camera motion in this setting. SfM is rarely used by itself to

localize a moving camera, and as such, global optimization in the form of bundle adjustment is generally adopted as the final step [12]. Mouragnon *et. al.* [14] present an approach to localize a single moving camera with a local bundle adjustment method to enable real time operation. With a single camera, however, 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 beacons 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. Nister *et. al.* [15] present an algorithm for visual odometry where stereo camera based SfM methods are used to localize the moving stereo platform. Agarwal and Konolige [16], and Oskiper *et. al.* [3] also present stereo based approaches to localization. Se *et. al.* [17] present a three camera based stereo system that triangulates SIFT feature correspondences between the cameras to localize a robot mounted with the camera rig. A laser scanner with a single camera can also be used to recover scale in translation estimates. Newman *et. al.* [18] present a system that uses a camera and a 3D laser scanner to localize a vehicle outdoors. In their system, SfM methods are used to obtain pose estimates, which initialize a scan matching algorithm for further refinement. They also globally correct their pose estimates using a loop closure scheme. The depths of all visual features are provided by their 3D laser scanner, resulting in the removal of the scale ambiguity for the translation parameters.

In this paper, we propose a new localization algorithm that integrates single camera SfM, and laser scan matching to localize a camera and 2D horizontal laser scanner mounted on a moving platform. Although 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. The 3D structure of the scene, however, is lost when it is projected onto the image plane. In addition, SfM techniques perform poorly when there are few, or no visual features in the images. We show that fusing the two sensors is likely to overcome some of the above shortcomings in order to improve localization accuracy.

We begin by introducing an Image Augmented Scan Matching (IASM) method that uses the color and texture cues around laser scan projections on images to determine scan point correspondences during the scan matching process. This method, however, can fail in situations with large viewpoint changes, or scenes with repeating patterns. To address such shortcomings in IASM, we develop a Homography based Image Augmented Scan Matching (H-IASM) approach. Specifically, we use SfM to determine the camera pose between successive images, which in turn, determines the homography mappings between the planes in the scene scanned by the laser scanner. Each scene plane has a unique homography associated with it, which is used to determine scan correspondences across images. We then refine these correspondences by applying a window based search method in the image space. We show that such a hybrid approach to localization results in a loop closure error of about 18cm, or 0.3%, on a 60m loop traversal in the interior corridor of a building. 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. Mourikis and Roumeliotis [4] achieve a 0.3% loop closure error with a camera-IMU system in a car moving outdoors at much higher speeds than our indoor system.

This paper is organized as follows. In Section 2 we present our extrinsic calibration method to find the relative orientation between a 2D horizontal laser scanner and a camera. We introduce the IASM algorithm in Section 3, and discuss its performance on an indoor data set. In Section 4 we provide an overview of existing SfM methods for standard visual odometry with specific implementation details. Our H-IASM algorithm is described in Section 5. In Section 6, we test the H-IASM algorithm on a large indoor dataset and report on the performance of a Kalman filter based fusion approach. Conclusions and future research are presented in Section 7.

2. Extrinsic Sensor Calibration

The relative rigid transformation between the camera and the laser scanner is needed to effectively fuse the two sensors. While the laser scanner does not require any intrinsic calibration, 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. A laser scanner measures the depth to a 3D point in space directly, whereas a camera can only measure the vector, originating from its center, along which the 3D point lies. We employ space resection to determine the position of the points along the image vectors, recovering their coordinates with respect to the camera. Any three world points recovered in camera coordinates, along with their laser coordinates can then be used to determine the relative pose between the sensors. The procedure is shown in Fig. 1, and is explained as follows.

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,

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

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 vect-

or of the directional line from the camera center to $\boldsymbol{p}^c$ is,

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

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, $[\boldsymbol{P_1}, \boldsymbol{P_2}, \boldsymbol{P_3}]$, as shown in Fig.1. These pairs are used by the 3-point algorithm [20] to determine the distance to the world points from the camera center, thus recovering their position in camera coordinates. The relative pose between the sensors is now obtained by applying Horn's method [21] to the three point pairs in laser and recovered camera coordinates.

Algorithm 1 describes our proposed calibration procedure, and is summarized as follows. We use a thin rectangular box placed at the height of the laser as the calibration target, as shown in Fig. 2(a). Scans and images of the target are collected from different locations by moving the sensor platform. In step 1 of Algorithm 1, lines are extracted from the laser scans using a combined least squares region growing method described later in Section 3.2. In step 2, we manually select the line in the laser scan corresponding to a side of the rectangular calibration box as shown in Fig. 2(b), and choose its endpoints as laser coordinates of the 3D points. We also manually select the image pixel locations of the endpoints of the calibration target in the corresponding camera image, shown in Fig 2(a). This process is repeated for all N locations of the sensor platform. Next, we apply the 3-point algorithm and Horn's method to recover the pose between the laser scanner and camera. In practice we collect more than 3 endpoint sets from different locations in order to obtain a more robust solution within a RANSAC framework as described in steps 3 to 8. The final solution is refined with a stage of Levenberg-Marquardt optimization in step 9.

To project laser scans onto images, we first transform each scan point $\boldsymbol{p}^l$ to the camera coordinate frame as,

$$\boldsymbol{p}^c = \boldsymbol{R}_l^c \boldsymbol{p}^l + \boldsymbol{t}_l^c \quad (3)$$

where, $[\boldsymbol{R}_l^c, \boldsymbol{t}_l^c]$ are the estimated rotation and translation from laser to camera frame of reference. We then find the image coordinates of the point using Eqn. (1). Fig. 2(a) shows the projection of a scan onto its corresponding image with the computed extrinsic calibration.

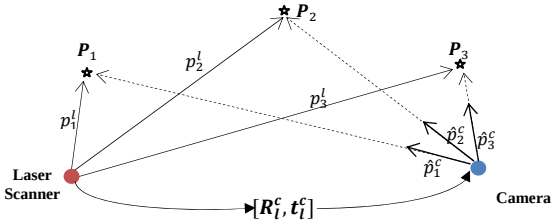


Figure 1: Three pairs of laser-camera correspondences serve to determine the relative sensor pose.

Algorithm 1: Laser - Camera Extrinsic Calibration

Input: Laser Scans $\boldsymbol{L}_{1..N}$ and Images $\boldsymbol{I}_{1..N}$ of calibration target
Output: Relative Pose $[\boldsymbol{R}_l^c, \boldsymbol{t}_l^c]$ between Laser and Camera.

1: Extract lines in laser scans.
2: Find both end points of rectangular box target in all scan lines, $\{x^l, y^l, 0\}_{1..2N}$, and corresponding images, $\{p_x, p_y, 1\}_{1..2N}$ for all N sensor locations..
3: FOR M RANSAC iterations
4: Randomly choose 3 scan point-image pixel pairs from step 2.
5: Use 3-point algorithm [20] to find unknown position of candidate scan points in camera coordinates.
6. Use Horn's method [21] to find relative pose between the 3 scan points chosen in step 4 and the corresponding 3 camera points determined in step 5.
7: Reproject all laser points onto corresponding images with pose found in step 6 and compute reprojection error.
8: IF error is below a threshold BREAK.
9: Refine winning pose solution with a Levenberg-Marquardt optimization step.

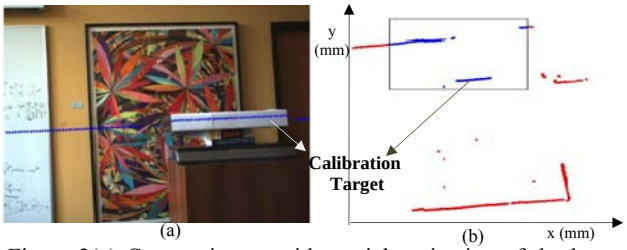


Figure-2(a) Camera image with partial projection of the laser scan shown in blue.(b) Corresponding laser scan with superimposed rectangular camera frame. The scan points overlapping camera's field of view are in blue, and those outside are in red.

3. Image Augmented Scan Matching (IASM)

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 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.,

$$\min_{\boldsymbol{R}_l, \boldsymbol{t}_l} \sum_{\mathrm{N}} \|\boldsymbol{m}_i - \boldsymbol{R}_l \boldsymbol{d}_i - \boldsymbol{t}_l\|^2 . \quad (4)$$

A naïve nearest neighbor approach to find point correspondences fails when the environment being scanned has minimal geometric features. Our approach in IASM is to exploit the rich color and texture information of camera images to find laser point correspondences. We assign scan point correspondences across successive scans by using the visual information around each scan point projection on the images. We use these correspondences to compute the transformation between the two successive scans within a RANSAC framework. The hypothesis evaluation scheme used in our RANSAC routine depends on

the distribution of planes in the scene. Specifically, for environments with rich geometric features, we use a scan alignment based evaluation metric, while for those with poor geometric features, we employ an image based metric. The details of the IASM algorithm are provided in the remainder of this section.

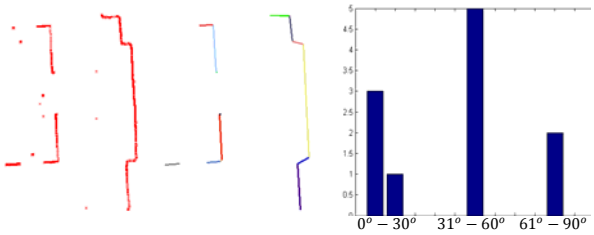
3.1. Image Based Nearest Neighbor Search

Laser scans of a scene from two different locations are projected onto their corresponding images. The scan projection tracker finds the best scan point correspondences across the two images as described below.

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. 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).

3.2. Line Extraction

In order to determine the richness of geometric features in the scene, we first apply a combined least squares-region growing method and extract lines from the laser scans. Since the scan points from the horizontal laser scanner are sorted in angular space, the first two points are chosen to form a seed line. The distance of the next scan point to this seed line is computed, and if it is greater than a threshold, a new line is initialized. Otherwise, a least squares step determines the line with this new point included. If this line has a similar slope-intercept form as the previous line, then the point is assigned to this growing line segment. We have set the similarity threshold to a sufficiently large value so as to accommodate sensor noise. Whenever the line points in a different direction, the growing segment is stopped, and a new line is started. This procedure efficiently estimates the lines in a scan, and also serves to filter noisy point returns. Fig. 3(b) is an example of the lines extracted from the scan in Fig. 3(a).



(a) (b) (c)

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

3.3. 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 distribution of filled angular bins is sufficiently wide, then a laser based metric to evaluate the RANSAC hypothesis is instantiated. 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 shown in Fig. 3(c).

On the other hand, if only one or two neighboring angle histogram bins are filled, then an image based evaluation method is used. 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 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.

Even though it might appear that applying an image based metric to determine the pose hypothesis score for all scans is the simplest approach, it could result in pose accumulation errors over time due to small extrinsic calibration errors. In addition, the laser scan might project on parts of the camera image with insufficient features and texture, resulting in inaccurate matching. In essence, the scan alignment metric prevents such error accumulation when there is a wide distribution of distinct geometric features in the scene.

3.4. IASM Experimental Results

A sample data set has been collected in a cubicle environment with narrow corridors by mounting the laser-camera rig on a moving platform. The navigation results are shown in Fig. 4. The loop closure error of IASM on a 35m loop traversal is 95cm, while the error in the ICP estimate is on the order of several meters.

The image patch based matching process described in this section fails if significant viewpoint changes occur between two images. This generally occurs while turning

the sensor platform, or when the sensors are very close to the walls of a hallway. Repeating patterns in the images also tend to result in incorrect point correspondences, as shown in Fig. 5(a). Thus, a robust scan point correspondence algorithm to handle such scenarios is needed. As shown in Section 6, determining the epipolar constraints between successive images can filter out incorrect scan point matches, and improve the scan point correspondence accuracy. In the following section we discuss our approach to determine the epipolar geometry between images.

Figure-4: Localization of sensor platform in an office environment. For a 35m loop, a loop closure error of IASM, shown in red, is 95 cm, while ICP, shown in green, fails to localize accurately.

4. Visual Odometry

Image sequences from a camera provide sufficient information to determine a camera's trajectory. In the SfM 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,

$$(\boldsymbol{K}^{-1}\boldsymbol{p}')^T\boldsymbol{E}(\boldsymbol{K}^{-1}\boldsymbol{p}) = 0 \qquad (5)$$

The 5-Point algorithm [13] determines the essential matrix in scenes with planar degeneracies which are ubiquitous in indoor environments. As the name suggests the algorithm determines $\boldsymbol{E}$ given 5 image feature correspondences. The epipolar geometry computation is in general most precise when sufficient motion occurs between two image frames. Hence, we choose to detect and track SIFT features [22] across successive images until the number of correspondences falls below a preset threshold. 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 iterative refinement to polish the winning hypothesis.

The convenient structure of the essential matrix $\boldsymbol{E}$ allows it to be decomposed into a rotation and translation because, $\boldsymbol{E} = [\hat{\boldsymbol{t}}_c]_\times \boldsymbol{R}_c$, where $[\boldsymbol{R}_c, \hat{\boldsymbol{t}}_c]$ represent the camera rotation and unit translation direction. We recover the scale of the translation, s, as follows. The 3D coordinates of a single point, $\boldsymbol{P}$, and its location in the first and last image in the tracked image sequence are obtained from the IASM process of Section 3. 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., $\hat{\boldsymbol{P}}$. The scale in the translation is then obtained by dividing the actual distance to the point as obtained by the laser scanner, by the triangulated distance, i.e.,

$$s = \|\boldsymbol{P}\| / \|\hat{\boldsymbol{P}}\| \qquad (6)$$

The triangulation procedure adopted is described in detail in [13].

5. Homography based Matching (H-IASM)

The basic IASM algorithm discussed in Section 3 is a brute force approach to finding scan point correspondences since it computes the SAD of large image patches around all scan point projections in successive images. This approach has problems in situations with large rotations and translations, or in scenes with repeating patterns as shown in Fig. 5(a). Furthermore, the scan points of the 2D horizontal laser scanner might project onto parts of the image with limited texture and features. Thus, there is a need to improve upon the basic IASM approach introduced earlier.

In general, indoor environments contain many planar regions such as walls. The mapping between the image projections of a scene plane viewed from two different positions can be represented by a linear transformation referred to as planar Homography, $\boldsymbol{H}$. A horizontal laser scanner captures these planar regions as lines. Therefore, the mapping between successive scan lines can be estimated by determining the homographies of the planes corresponding to each scan line.

At the outset, we assume that the laser scan plane is approximately perpendicular to the scene planes. We transform each laser scan to camera coordinates, and extract lines from it as described in Section 3.2.Line AB in Fig. 6. corresponds to one such line, lying on the vertical scene plane, Π, which is assumed to be perpendicular to the horizontal scanning plane, SAB. We then compute the unit normal, $\hat{\boldsymbol{n}}$, to AB in the scan plane, SAB. The perpendicular distance, d, between the camera center, C, and the scene plane, Π, is obtained as the dot product of the normal $\hat{\boldsymbol{n}}$ with any point $\boldsymbol{P}_i$ on the scan line in the camera coordinates, i.e.,

$$d = \hat{\boldsymbol{n}}^T\boldsymbol{P}_i, \quad \text{or} \quad \hat{\boldsymbol{n}}^T\boldsymbol{P}_i/d = 1 \qquad (7)$$

The corresponding location of point $\boldsymbol{P}_i$ in the second view, $\boldsymbol{P}_i'$,is given by,

$$\boldsymbol{P}_i' = \boldsymbol{R}_c\boldsymbol{P}_i + s_t\hat{\boldsymbol{t}}_c \qquad (8)$$

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. Since this scale is unknown prior to finding scan point correspondences in images, we opt to use the scale computed in the previous

iteration, s_{t-1}, instead. Substituting Eqn. (7) in (8) and reordering the terms,

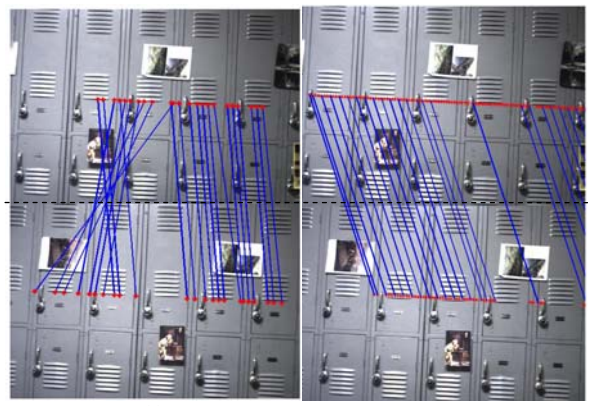
$$\boldsymbol{P}_i' = \left(\boldsymbol{R}_c + \frac{s_{t-1}}{d}\hat{\boldsymbol{t}}_c\hat{\boldsymbol{n}}^T\right)\boldsymbol{P}_i = \boldsymbol{H}\boldsymbol{P}_i \quad (9)$$

where,

$$\boldsymbol{H} = \boldsymbol{R}_c + \frac{s_{t-1}}{d}\hat{\boldsymbol{t}}_c\hat{\boldsymbol{n}}^T. \quad (10)$$

$\boldsymbol{H}$ represents the approximate homography that maps a scan point $\boldsymbol{P}_i$ in the first view to $\boldsymbol{P}_i'$ in the second view. Obtaining an accurate mapping using the above approach is rarely possible as the homography is a function of the scale computed in the previous iteration. Furthermore, while computing the normal, $\hat{\boldsymbol{n}}$, to any plane in the scene, it is assumed that the scene planes are orthogonal to the scan plane which in practice might not always be true. Hence, we choose to use the computed homography as an approximate transformation that limits the search area for IASM. We have empirically found this homography approach to improve the accuracy and robustness of the scan matching process in indoor environments with planar surfaces. Fig. 5(b) depicts a scene with repeating textures where H-IASM finds scan point correspondences accurately, while the IASM approach fails, as shown in Fig. 5(a).

Fig. 7 shows the flowchart of the H-IASM 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 H-IASM sub-systems. The visual odometry system computes the rotation and unit translation. In the H-IASM system, lines are extracted from the laser scan, which along with the camera pose matrices from visual odometry, are used to compute a homography for each plane in the scan as described earlier.



(a) (b)

Figure-5: (a) Incorrect scan point correspondences using IASM in images with repeating patterns; (b) accurate matches obtained using H-IASM by computing homography between scan lines of planes in images.

These homographies provide a local search region to find scan point correspondences in the images. The patch based search method described in Section 3.1 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 window defined by each homography mapping. Once correspondences are found, the robust RANSAC based method described in Section 3.3 determines the pose transformation. The scale in the current translation is also computed to be used with the image-scan pair in the subsequent iteration. Since scale is unavailable during initialization, the basic IASM method of Section 3 is used with the first image and scan pair.

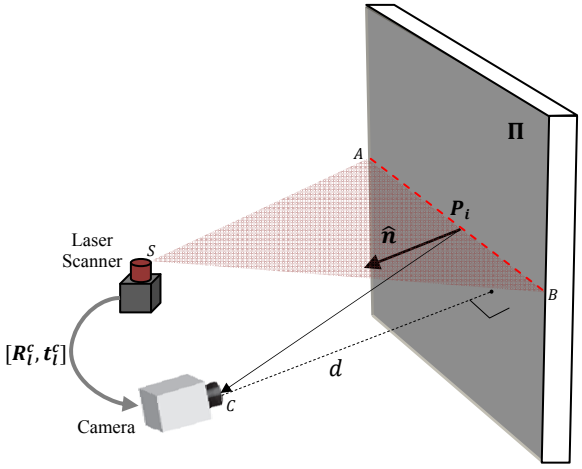


Figure-6: The normal, $\hat{\boldsymbol{n}}$, to the scan line, AB, is computed in the laser scanning plane, SAB, in camera coordinates. The perpendicular distance between the scene plane, **Π**, and the camera center, C, is d.

6. Results

We compare the accuracy of the H-IASM algorithm presented in this paper with the ground truth, that has been collected using an Applanix position and orientation system used for land surveying. This is an aided inertial navigation system which consists 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 0.003 deg/hr, and three precision accelerometers with bias of less than 25µg. 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-updates (ZUPTs), which allow for accumulated biases in the IMU to be estimated, and any velocity drift to be corrected. These ZUPT points manifest as discontinuity points in the ground truth paths of Fig. 8 to be discussed shortly. In our tests, the IMU system used for ground truth has a loop closure error of approximately 3 cm, i.e. 0.05% for a 60m loop. The ground truth was collected at a different time from the data set, as the Honeywell IMU was unavailable

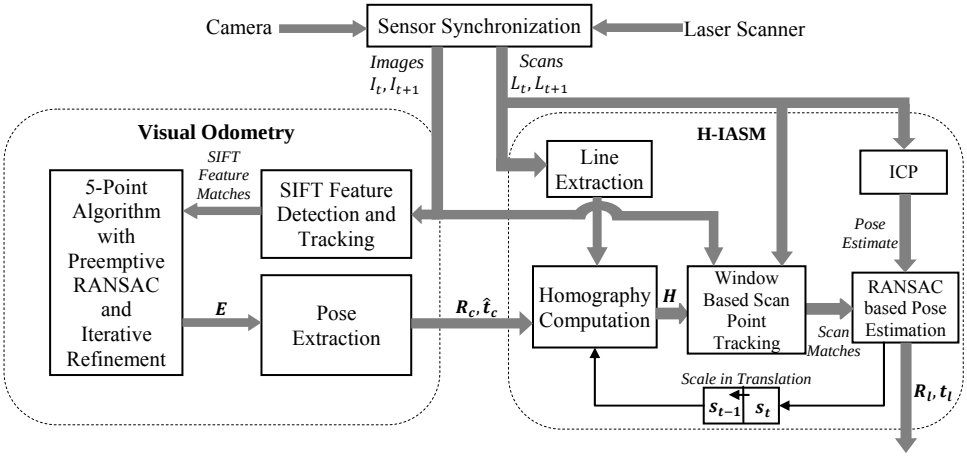


Figure-7: Flow diagram of H-IASM.

at the time of data collection. Thus, waypoints were set by comparing camera images obtained during data capture and those during ground truth collection. The start, stop, and turn points however, were replicated precisely while traversing the hallway the second time.

Ground truth comparisons of H-IASM and IASM for a 60m loop are shown in Figs. 8(a) and 8(b) respectively. The raw visual odometry and ICP results are plotted against ground truth in Fig. 8(c). The loop closure error for various schemes are shown in Table 1. As expected, the loop closure error is the lowest for H-IASM at 18cm. Specifically, the loop closure error for H-IASM is a factor of 13 smaller than ICP, a factor of 5 smaller than visual odometry, and a factor of 6 smaller than IASM.

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. As seen in Table 1, the average position error for H-IASM is about a factor of 50 smaller than ICP, a factor of 6 smaller than visual odometry, and a factor of 2.5 smaller than IASM.

To compare our approach with traditional sensor fusion approaches, the visual odometry and ICP based odometry pose estimates are fused within a Kalman Filter (KF) framework. ICP pose estimates are in two dimensions, while the visual odometry estimates are in three. Thus, the latter is reduced to two dimensions by extracting the yaw angle from the 3D rotation matrix, and dropping the third dimension in the translation estimate. The state vector, $\boldsymbol{S}$, adopted in the KF is six-dimensional, consisting of planar translation and rotation in addition to velocities.

$$\boldsymbol{S} = [x, y, \omega, \dot{x}, \dot{y}, \dot{\omega}] \qquad (11)$$

The process model is that of a constant-velocity particle. Each estimate is treated as a black box input with fixed measurement covariance estimates. The incremental rotation and translation estimates from each method are converted to angular and linear velocity estimates before being input to the filter. The measurement covariances have been set such that forward translation estimates from ICP are given less weight than that of visual odometry, and the 2D rotation estimates of ICP are given a higher weight.

Fig. 8(d) depicts the results of Kalman filtering the visual odometry and ICP pose estimates. The loop closure error and average position error for the KF approach are also shown in Table 1. Even though KF improves upon ICP or visual odometry alone, its performance is inferior to that of H-IASM.

7. Conclusions and Future Work

In this paper, a novel image augmented laser scan matching algorithm has been presented for indoor navigation, resulting in a 0.3% loop closure error. This is better than the loop closure error obtained in [3] for a combined indoor-outdoor path with a more elaborate system made of four cameras and an IMU. Future work involves bundle adjustment and automatic loop closure detection. We believe that the increased localization accuracy of our approach can potentially increase the accuracy of detecting loop closures. Ultimately, we plan on applying our proposed algorithm to localize a backpack mounted with laser scanners and cameras for 3D indoor modeling.

Localization Method	Loop Closure Error (m)	Average Position Error (m)
ICP	2.49	6.99
VO	0.88	0.81
KF	0.57	0.37
IASM	1.14	0.35
H-IASM	**0.18**	**0.14**

Table-1: A comparison of the mean position and loop closure errors for ICP, Visual Odometry (VO), Kalman filtered path, IASM, and H-IASM.

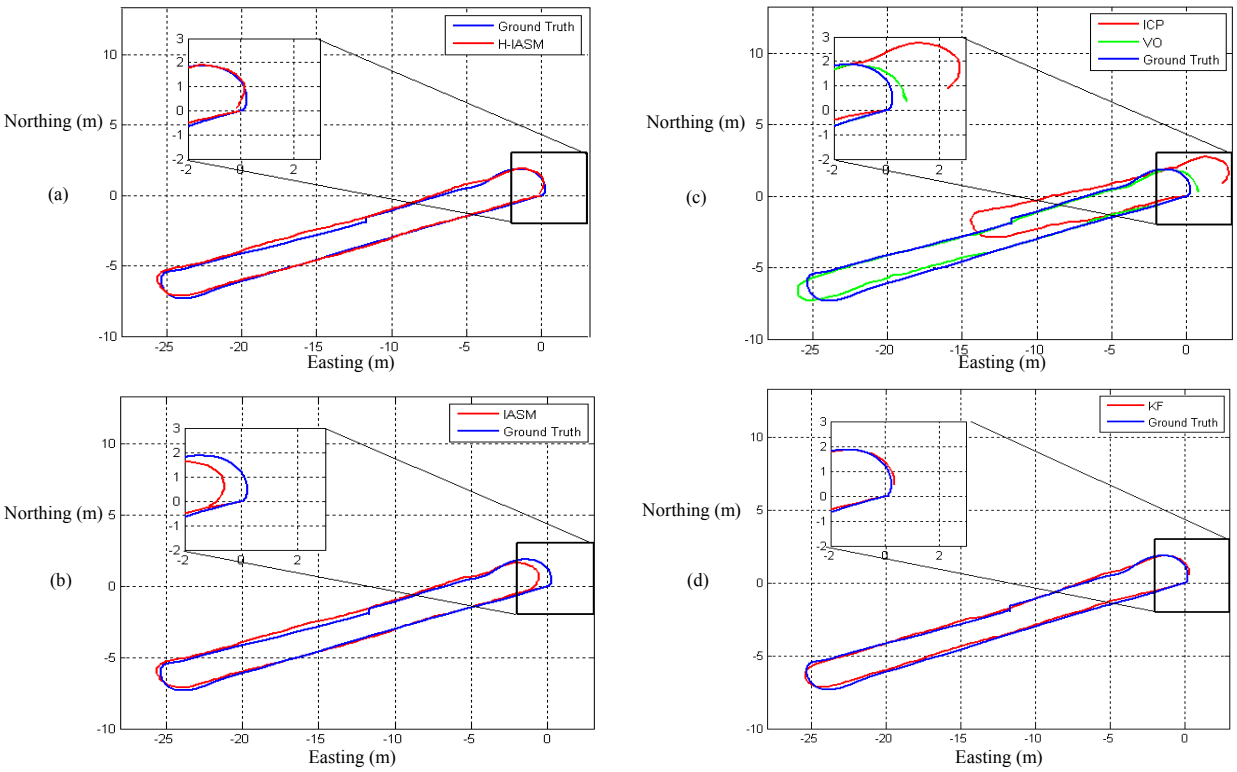


Figure-8:(a) Reconstructed H-IASM path, in red, and ground truth in blue. (b) Reconstructed IASM path, in red, and ground truth in blue. (c) The raw ICP path, in red, and visual odometry path, in green against ground truth, in blue. (d) The path obtained by applying a Kalman Filter to raw ICP and visual odometry, shown in red, compared against ground truth, in blue.

8. 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.

[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, *FastSLAM 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.

[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*, International Journal of Computer Vision, 2004.