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 scan correspondence estimates which are refined on the images. We have demonstrated our proposed system, consisting of a laser scanner and a camera, to result in a 0.4% loop closure error for a 60m loop around the interior corridor of a building.

I. 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 objects of known size are placed in the

This research project is supported by the ARO grant W911NF-07-1-0471 and the HSN MURI grant W911NF-06-1-0076.
## Image Augmented Laser Scan Matching for Indoor Localization

### Abstract

Approved for public release; distribution unlimited

### Subject Terms

<table>
<thead>
<tr>
<th>a. REPORT</th>
<th>b. ABSTRACT</th>
<th>c. THIS PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>unclassified</td>
<td>unclassified</td>
<td>unclassified</td>
</tr>
</tbody>
</table>

### Security Classification of:

- a. REPORT: unclassified
- b. ABSTRACT: unclassified
- c. THIS PAGE: unclassified

### Limitation of Abstract

Same as Report (SAR)

### Number of Pages

7
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, visual odometry 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 visual odometry, 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, visual odometry 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 Visual Odometry based Scan Matching (VOSM) approach. Specifically, we use visual odometry to determine the camera pose between successive images, which in turn aids in determining scan correspondence estimates across images. We then refine these correspondence estimates 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 25cm, or 0.4%, 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 II 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 III, and discuss its performance on an indoor data set. In Section IV we provide an overview of existing visual odometry methods with specific implementation details. Our VOSM algorithm is described in Section V. In Section VI, we test the VOSM 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 VII.

II. 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, \( p^c = [x^c, y^c, z^c]^T \), is represented in image coordinates as,

\[
p = \begin{bmatrix} p_x & p_y & 1 \end{bmatrix}^T = K \begin{bmatrix} x^c / z^c & y^c / z^c & 1 \end{bmatrix}^T
\]

(1)

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

\[
\vec{p} = K^{-1} p / \|K^{-1} p\|
\]

(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 \( p^l = [x^l, y^l, 0]^T \). We begin by manually choosing three (laser point, image vector) pairs, i.e., \( ([p^l_1, p^l_2, p^l_3] \leftrightarrow [\vec{p}^c_1, \vec{p}^c_2, \vec{p}^c_3]) \), corresponding to three world points, \( \{P_1, P_2, P_3\} \), 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 subsequently. 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 \( p^l \) to the camera coordinate frame as,

\[
p^c = R^c_t p^l + t^c \tag{3}
\]

where, \([R^c_t, t^c] \) 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.

**Algorithm 1**: Laser - Camera Extrinsic Calibration

**Input**: Laser Scans \( L_{1,N} \) and Images \( I_{1,N} \) of calibration target

**Output**: Relative Pose \([R^c_t, t^c] \) 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.

**III. 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, \([R^s_t, t^s] \), by minimizing the squared distance between each of the N points in the first scan, \( m_i \), and their nearest neighbors in the second scan, \( d_i \), i.e.,

\[
\min_{R^s_t, t^s} \sum_{i=1}^{N} \| m_i - R^s_t d_i - t_i \|^2. \tag{4}
\]

A naïve nearest neighbor approach to find point correspondences fails when the environment being scanned has few 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.

A. 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, \([L_t, L_{t+1}]\), are projected onto their corresponding images, \([I_t, I_{t+1}]\).
2. Image patches are extracted around each scan point projection in images \(I_t\) and \(I_{t+1}\) in order to find patch correspondences across images by minimizing the bidirectional Sum of Absolute Difference (SAD).

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.

C. 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 95 cm, 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 VI, 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.
are enforced by the essential matrix, $E$, such that, for any two calibrated point correspondences $p \leftrightarrow p'$, we have,

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

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 $E$ 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 $E$ allows it to be decomposed into a rotation and translation because, $E = [\hat{e}_c]_s R_c$, where $[R_c, \hat{e}_c]$ 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, $P$, and its location in the first and last image in the tracked image sequence are obtained from the IASM process of Section III. This image correspondence pair is triangulated with the current camera pose estimate, $[R_c, \hat{e}_c]$, to determine the scaled coordinates of the point, i.e., $\hat{P}$. 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.,

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

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

![Fig. 5. (a) Incorrect scan point correspondences using IASM in images with repeating patterns; (b) accurate matches obtained using VOSM by computing homography between scan lines of planes in images.](image)

**V. VISUAL ODOMETRY BASED SCAN MATCHING (VOSM)**

The basic IASM algorithm discussed in Section III 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.

The corresponding location in the second view, $P'_s$, of a point $P$ in the first view is given by,

$$P'_s = R_c P + s \hat{e}_c \quad (7)$$

where, $[R_c, \hat{e}_c]$, are the camera rotation and unit translation obtained from visual odometry, and $s$ 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. The location of this point in the second image can then be computed using eqn. 1.

Obtaining an accurate mapping using the above approach is rarely possible as the point projection is a function of the scale computed in the previous iteration. Hence, we choose to use the computed transformation as an approximate image projection that limits the search area for IASM. We have empirically found this approach to improve the accuracy and robustness of the scan matching process in indoor environments. Fig. 5(b) depicts a scene with repeating textures where VOSM finds scan point correspondences accurately, while the IASM approach fails, as shown in Fig. 5(a).

Fig. 6 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 IASM sub-systems. The visual odometry system computes the rotation and unit translation. With the scale computed in the previous iterations, the camera pose matrix is used to transform the laser points in the first image's coordinate system to the second image's coordinates. Projecting these transformed laser points on to the 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 III-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 III-B 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 III is used with the first image and scan pair.

**VI. RESULTS**

We compare the accuracy of the VOSM 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 had loop closure errors of approximately 3 cm, i.e. 0.05% for a 60m loop.

Ground truth comparisons of VOSM and IASM for a 60m loop are shown in Figs. 7(a) and 7(b) respectively. The raw visual odometry and ICP results are plotted against ground truth in Fig. 7(c). The loop closure error for various schemes are shown in Table 1. As expected, the loop closure error is the lowest for VOSM at 25cm. Specifically, the loop closure error for VOSM is a factor of 18 smaller than ICP, a factor of 3 smaller than visual odometry, and a factor of 2 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 VOSM is about a factor of 37 smaller than ICP, a factor of 14 smaller than visual odometry, and a factor of 6 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, $\mathbf{S}$, adopted in the KF is six-dimensional, consisting of planar translation and rotation in addition to velocities.

$$\mathbf{S} = [x, y, \omega, \dot{x}, \dot{y}, \dot{\omega}]$$ (8)

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 and 2D rotation estimates from ICP are given less weight than that of visual odometry. Further, the visual odometry estimates are found to be poor in cross-track translation, for which the covariance has been set to a larger value than that of ICP.

Fig. 7(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 IASM and VOSM.

VII. CONCLUSIONS AND FUTURE WORK

In this paper, a novel image augmented laser scan matching algorithm has been presented for indoor navigation, resulting in a 0.4% 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.

<table>
<thead>
<tr>
<th>Localization Method</th>
<th>Loop Closure Error (m)</th>
<th>Average Position Error (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICP</td>
<td>4.26</td>
<td>5.61</td>
</tr>
<tr>
<td>VO</td>
<td>0.69</td>
<td>2.10</td>
</tr>
<tr>
<td>KF</td>
<td>0.64</td>
<td>1.75</td>
</tr>
<tr>
<td>IASM</td>
<td>0.43</td>
<td>0.85</td>
</tr>
<tr>
<td>VOSM</td>
<td><strong>0.25</strong></td>
<td><strong>0.15</strong></td>
</tr>
</tbody>
</table>

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


