Toward Mutual Information based Automatic Registration of 3D Point Clouds

Gaurav Pandey, James R. McBride, Silvio Savarese and Ryan M. Eustice

***Abstract*— This paper reports a novel mutual information (MI) based algorithm for automatic registration of unstructured 3D point clouds comprised of co-registered 3D lidar and camera imagery. The proposed method provides a robust and principled framework for fusing the complementary information obtained from these two different sensing modalities. High-dimensional features are extracted from a training set of textured point clouds (*scans*) and hierarchical k-means clustering is used to quantize these features into a set of *codewords*. Using this codebook, any new scan can be represented as a collection of codewords. Under the correct rigid-body transformation aligning two overlapping scans, the MI between the codewords present in the scans is maximized. We apply a James-Stein-type shrinkage estimator to estimate the true MI from the marginal and joint histograms of the codewords extracted from the scans. Experimental results using scans obtained by a vehicle equipped with a 3D laser scanner and an omnidirectional camera are used to validate the robustness of the proposed algorithm over a wide range of initial conditions. We also show that the proposed method works well with 3D data alone.**

I. Introduction

A fundamental requirement of mobile robots is to sense and understand the environment around them. Two important categories of perception sensors typically used are: *(i)* range sensors (e.g., 3D/2D lidars, radars, sonars) and *(ii)* cameras (e.g., perspective, stereo, omnidirectional). To create realistic 3D maps from these sensors requires precisely aligning camera data onto range information and vice versa. To accomplish this task, camera and range sensors must be extrinsically calibrated [1]–[3], which allows for the association of the two modalities intra-scan; however, creating full 3D models of a large environment requires the automatic inter-scan alignment of hundreds or thousands of scans.

One of the most common methods of scan alignment is iterative closest point (ICP) and was first introduced by Besl and McKay [4]. In their work, they proposed a method to minimize the Euclidean distance between corresponding points to obtain the relative transformation between two scans. Chen and Medioni [5] further introduced the point-to-plane variant of ICP owing to the fact that most of the range measurements are typically sampled from a locally planar surface. Similarly, Alshawa [6] introduced a line-based matching variant called iterative closest line (ICL). In ICL, the line features are extracted from the range scans and aligned to obtain the rigid-body transformation. Several other variants of the ICP algorithm have also been proposed and can be found in the survey paper by Rusinkiewicz and Levoy [7].

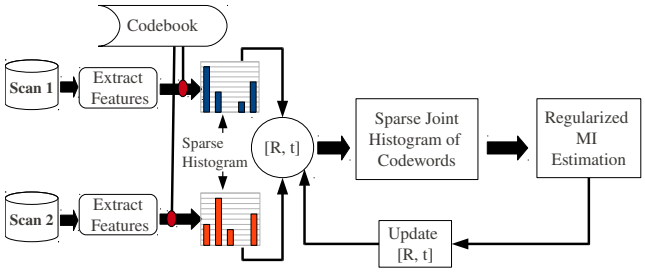


Fig. 1. The proposed scan registration method learns a codebook of the high-dimensional features extracted from the scans. Using this codebook the empirical histograms of codewords present in the scans are computed for a given rigid-body transformation. The MI is optimally estimated from them using a James-Stein-type shrinkage estimator. This MI is maximized at the optimal transformation parameters that aligns the two scans.

One of the main reasons for the popularity of ICP-based methods is that it solely depends on the 3D points and does not require extraction of complex geometric primitives. Moreover, the speed of the algorithm is greatly boosted when it is implemented with kd-trees for establishing point correspondences. However, most of the deterministic algorithms discussed so far do not account for the fact that in real-world datasets, when the scans are coming from two different time instances, we never achieve exact point correspondence. Moreover, scans are generally only partially overlapped—making it hard to establish point correspondences by applying a threshold on the point-to-point distance.

Recently, several probabilistic techniques have been proposed that model the real-world data better than the deterministic methods. Biber et al. [8] apply a probabilistic model by assuming that the second scan is generated from the first through a random process. Haehnel and Burgard [9] apply ray-tracing techniques to maximize the probability of alignment. Biber [10] also introduced an alternate representation of range scans, the normal distribution transform (NDT), where he subdivides a 2D plane into cells and assigns a normal distribution to each cell to model the distribution of points in that cell. This density is used to match scans and, therefore, no explicit point correspondence is required. Segal

*This work was supported by Ford Motor Company via the Ford-UofM Alliance.

G. Pandey and S. Savarese are with the Department of Electrical Engineering & Computer Science, University of Michigan, Ann Arbor, MI 48109, USA {pgaurav,silvio}@umich.edu

J. McBride is with the Research & Innovation Center, Ford Motor Company, Dearborn, MI 48124, USA jmcbride@ford.com

R. Eustice is with the Department of Naval Architecture & Marine Engineering, University of Michigan, Ann Arbor, MI 48109, USA eustice@umich.edu

et al. [11] proposed to combine the iterative closest point and point-to-plane ICP algorithms into a single probabilistic framework called generalized ICP (GICP). They devised a generalized framework that naturally converges to point-to-point or point-to-plane ICP by appropriately defining the sample covariance matrices associated with each point. Their method exploits the locally planar structure of both participating scans as opposed to just a single scan as in the case of point-to-plane ICP. They have shown promising results with full 3D scans acquired from a Velodyne laser scanner.

Most of the ICP algorithms described above are based on 3D point clouds alone and very few incorporate visual information into the ICP framework. Johnson and Kang [12] proposed a simple approach to incorporating color information into the ICP framework by augmenting the three color channels to the 3D coordinates of the point cloud. Akca et al. [13] proposed a novel method of using intensity information for scan matching. They proposed the concept of a quasi-surface, which is generated by scaling the normal at a given 3D point by its color, and then matching the geometrical surface and the quasi-surfaces in a combined estimation model. This approach works well when the environment is structured and the normals are well defined. Pandey et al. [14] proposed an algorithm for bootstrapping the ICP algorithm using camera data. They exploit the co-registration of the 3D point cloud with the available camera imagery to associate high-dimensional feature descriptors such as scale invariant feature transform (SIFT) [15] or speeded up robust features (SURF) [16] to the 3D points. They first establish putative point correspondence in the high-dimensional feature space and then use these correspondences in a random sample consensus (RANSAC) framework to obtain an initial rigid-body transformation that aligns the two scans. This initial transformation is then refined in a GICP [11] framework.

All of the aforementioned methods either use the point cloud data alone or use the data from the two modalities (camera/lidar) in a decoupled way, without exploiting the statistical dependence of the multi-modal data. It is important to note that the camera image and the lidar point cloud are statistically dependent on each other; because, the underlying structure generating the two signals (3D point cloud / image) is the same. It is not new to fuse multi-modal data by exploiting their statistical dependence. In fact, registration of multi-modal data by maximizing the mutual information (MI) has been state-of-the-art in the medical imaging community for over a decade. The idea of MI-based multi-modal image registration was first introduced by Viola et al. [17] and Maes et al. [18]. Since then, researchers (especially in medical imaging) have widely used the MI framework to focus on specific registration problems in various clinical applications [19]. Within the robotics community, the application of MI has not been as widespread, even though robots today are often equipped with different modality sensors (e.g., camera/lidar). Here, we present a novel MI-based algorithm for automatic registration of unstructured 3D point clouds collected using co-registered 3D lidar and camera imagery

Fig. 2. Test vehicle (left). The 3D laser scanner and omnidirectional camera system mounted on the roof of the vehicle (right).

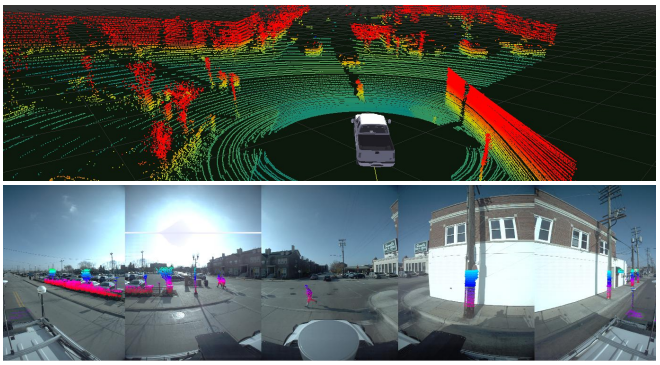


Fig. 3. The top panel is a perspective view of the Velodyne 3D lidar range data, color-coded by height above the ground plane. The bottom panel shows the above ground plane range data projected into the corresponding image from the Ladybug3 camera.

(Fig. 1). Our method provides a robust framework for incorporating complementary information obtained from these modalities into the registration process.

The remainder of this paper proceeds as follows: In Section II we describe the proposed method of automatic registration of 3D scans. In Section III we present results showing the robustness of the proposed method and present a comparison of our method with generalized ICP. Finally, in Section IV we summarize our findings.

II. Methodology

In our work we have used a Velodyne 3D laser scanner and a Point Grey Ladybug3 omnidirectional camera system mounted to the roof of a vehicle (Fig. 2). We assume that the intrinsic and extrinsic calibration parameters for these sensors are known. The calibration allows us to project 3D points from lidar onto the corresponding image (and vice versa) as depicted in Fig. 3. This co-registration allows us to extract high-dimensional feature descriptors from the image (SIFT [15], SURF [16], etc.) and associate them to a corresponding 3D lidar point that projects onto that pixel location. Moreover, we can also extract 3D features (fast point feature histogram (FPFH) [20], rotation invariant feature transform (RIFT) [21], spin-images [22], etc.) from the point cloud. We combine these features to form a robust high-dimensional feature vector, which can be calculated at some keypoints of the scan. This allows us to represent a scan as a collection of high-dimensional feature vectors. Thus, for any two overlapping scans the joint distribution of these features should show maximum correlation when viewed

(a) Sample images from training dataset (*ford campus*)

(b) Sample images from testing dataset (*downtown*)

Fig. 4. The codebook and target distribution are learned from the training dataset, and all experiments are performed on the testing dataset. It should be noted that the training and testing datasets are captured in similar outdoor urban environments, though not the same. It is important for the codebook to be representative, but the testing and training environments need not be identical.

under the correct rigid-body transformation. Here, we use concepts from statistics and information theory to formulate a MI-based cost function to solve the scan registration problem. An overview of the proposed method is shown in Fig. 1.

A. Theory

The mutual information between two random variables X and Y is a measure of their statistical dependence. Various formulations of MI are present in the literature, each of which demonstrate a measure of statistical dependence of the random variables in consideration. One such form of MI is defined in terms of entropy of the random variables:

$$\mathrm{MI}(X,Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y), \tag{1}$$

where $\mathrm{H}(X)$ and $\mathrm{H}(Y)$ are the entropies of random variables X and Y, respectively, and $\mathrm{H}(X,Y)$ is the joint entropy of the two random variables.

$$\mathrm{H}(X) = -\sum_{x\in X} p_X(x)\log p_X(x) \tag{2}$$

$$\mathrm{H}(Y) = -\sum_{y\in Y} p_Y(y)\log p_Y(y) \tag{3}$$

$$\mathrm{H}(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p_{XY}(x,y)\log p_{XY}(x,y) \tag{4}$$

The entropy $\mathrm{H}(X)$ of a random variable X denotes the amount of uncertainty in X, whereas $\mathrm{H}(X,Y)$ is the amount of uncertainty when the random variables X and Y are co-observed. Hence, (1) shows that $\mathrm{MI}(X,Y)$ is the reduction in the amount of uncertainty of the random variable X when we have some knowledge about random variable Y. In other words, $\mathrm{MI}(X,Y)$ is the amount of information that Y contains about X and vice versa.

B. Mathematical Formulation

We first create a dictionary of *codewords* representing the quantization of the high-dimensional features extracted in the scans. We extract N such features (training samples) from a set of scans called the training dataset (Fig. 4). We use a hierarchical k-means clustering [23] algorithm on the training samples to cluster the feature space into K clusters. The centroids of these clusters are defined as *codewords* $\{c_i; i = 1, 2, \cdots, K\}$ and the collection of these codewords is called the *codebook*. We use this codebook to map any feature vector to a unique integer i corresponding to the codeword c_i that gives a maximum similarity score with the feature vector.

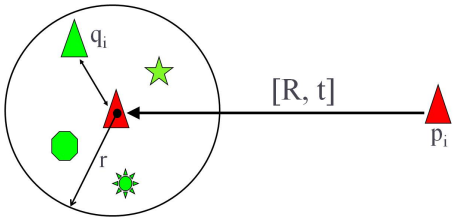


Fig. 5. Illustration of the nearest neighbour search algorithm used to establish codeword correspondence; each shape above represents a different codeword—green colorings belong to scan **Q** and red to scan **P**. The codeword c_i^q that gives the maximum similarity score with c_i^p is chosen as the correspondence.

We consider the collection of these codewords present in a scan as the random variables X and Y. The marginal and joint probabilities of these random variables, $p_X(x)$, $p_Y(y)$ and $p_{XY}(x,y)$, can be obtained from the normalized marginal and joint histograms of the codewords present in the scans that we want to align. Let **P** and **Q** be the two scans that we want to align. Let $C^P = \{c_i^p; i = 1, \cdots, n\}$ and $C^Q = \{c_i^q; i = 1, \cdots, m\}$ be the set of codewords, and $\{\mathbf{p}_i; i = 1, \cdots, n\}$ and $\{\mathbf{q}_i; i = 1, \cdots, m\}$ be the set of 3D points corresponding to the codewords present in scans **P** and **Q**, respectively. If the rigid-body transformation that perfectly aligns these scans is given by $[\mathrm{R}, \mathbf{t}]$, then the coordinate transformation of any point in scan **P** onto the reference frame of scan **Q** is given by:

$$\hat{\mathbf{q}}_i = \mathrm{R}\mathbf{p}_i + \mathbf{t}. \tag{5}$$

For a correct rigid-body transformation, the codeword c_i^p of point $\mathbf{p}_i$ should be the same as the codeword c_i^q of the corresponding point $\hat{\mathbf{q}}_i$. Thus, for a given rigid-body transformation, the corresponding codewords c_i^p and c_i^q are the observations of the random variables X and Y, respectively.

We use nearest neighbor search to establish the codeword correspondence (Fig. 5). A codeword c_i^p in scan **P** is first transformed to the reference frame of **Q**. All the codewords in scan **Q** that are within a sphere of radius r around c_i^p are considered as potential correspondences. The codeword c_i^q that gives the maximum similarity score with c_i^p is chosen as the correspondence. In the case where we have multiple codeword assignment within the sphere, then the codeword that is closest in Euclidean space to c_i^p takes precedence. We use this correspondence to create the joint histogram of codewords for the given transformation. The maximum likelihood estimate (MLE) of the marginal and joint probabilities of the random variables X and Y can be

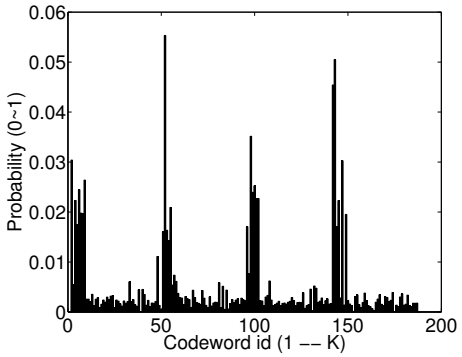
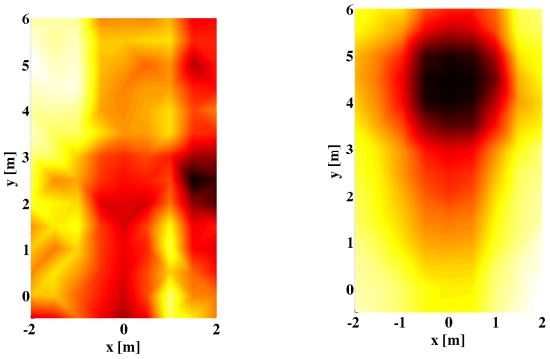


Fig. 6. A non-uniform target distribution estimated from the training dataset. The feature descriptors chosen here are a combination of FPFH and SURF extracted from the co-registered lidar and camera data. The size of the codebook is 200.



(a) Standard MI without James-Stein estimator (b) Shrinkage optimized MI with James-Stein estimator

Fig. 7. Top view of the MI cost-function surface versus the translation parameters x and y aligning the two scans. The correct value of translation is given by $(0.02, 4.31)$. Light to dark represents increasing values of the cost function.

obtained from the normalized marginal and joint histograms of these codewords.

It is important to note that the number of codewords extracted from a scan (i.e., n or m) is typically much less than the dimensions of the joint histogram (i.e., $K \times K$). Moreover, the number of different codewords present in any scan is generally only a fraction of the size of codebook. This causes most of the entries of the joint and marginal histograms to be unobserved, leading to high mean-squared-error (MSE) in the MLE due to overfitting. Therefore, we apply a James-Stein (JS) shrinkage approach to improve the MSE of the maximum likelihood (ML) estimator. This method was proposed by Hausser and Strimmer in [24] for entropy and MI estimation, and is based on shrinking the ML estimator of the distribution of a random variable Z toward a target distribution $\mathrm{T} = [\mathrm{T}_1, \mathrm{T}_2, \cdots, \mathrm{T}_K]$:

$$\hat{Z}_k^{JS} = \lambda \mathrm{T}_k + (1-\lambda)\hat{Z}_k^{ML}, \tag{6}$$

where $\hat{Z}_k = p_Z(z = k)$ and $\lambda \in [0, 1]$ is a shrinkage coefficient used to optimize the estimation of MI. The target distribution here refers to the distribution of codewords observed in an ideal case (i.e., when $n \gg K \times K$). If all the features extracted from a scan were equally likely, then a uniform distribution becomes an obvious choice for the target distribution. However, the occurrence of any feature extracted from the scans is dependent upon the environment. So, we learn the target distribution from the training dataset along with the codebook. The target distribution is estimated from the normalized histogram of codewords present in the training dataset. A sample target distribution corresponding to a particular codebook is shown in Fig. 6.

The shrinkage coefficient λ is calculated from the number of codewords obtained from nearest neighbor search:

$$\lambda = \frac{2}{(1 + \exp^{-c/\sigma})} - 1, \tag{7}$$

where c is the number of corresponding codewords and σ is a parameter proportional to the average number of codewords present in a scan. Thus, λ takes on a value between 0 (no correspondence / no shrinkage) and 1 (maximum correspondence / full shrinkage).

Once we have a good estimate of the joint and marginal probability distributions we can write the MI of the random variables (X, Y) as a function of the rigid-body transformation $[\mathrm{R}, \mathbf{t}]$, thereby formulating a cost function:

$$\mathbf{\Theta} = \underset{\mathbf{\Theta}}{\operatorname{argmax}}\, \mathrm{MI}(X, Y; \mathbf{\Theta}), \tag{8}$$

where $\mathbf{\Theta} = [x, y, z, \phi, \theta, \psi]^\top$ is the six degree of freedom (DOF) parametrization of the rigid-body transformation $[\mathrm{R}, \mathbf{t}]$.

The small number of codewords present in a scan make the estimation of MI a challenging task. The shrinkage approach described above provides a robust estimate of MI. Fig. 7 shows a comparison between the standard MI and the shrinkage optimized MI-based cost function. Clearly, the proposed shrinkage optimized MI-based cost function shows a global maxima at the desired rigid-body transformation. We use the simplex method proposed by Nelder and Mead [25] to estimate the optimum value of the registration parameter, $\mathbf{\Theta}$, that maximizes the cost function given in (8). The complete registration method is summarized in Algorithm 1.

III. Experiments and Results

We present results from real data collected from a 3D laser scanner (Velodyne HDL-64E) and an omnidirectional camera system (Point Grey Ladybug3) mounted on the roof of a Ford F-250 vehicle. We use the pose information available from a high end inertial measurement unit (IMU) (Applanix POS-LV 420 INS with Trimble GPS) as the ground-truth to compare the scan alignment errors. The datasets used in our experiments are available online [26] and are divided into two distinct runs: *(i) downtown* and *(ii) ford campus*, both taken in Dearborn, Michigan. We use the *downtown* dataset for testing and the *ford campus* dataset for learning the codebook and the target distribution. We performed the following experiments to analyze the robustness of the proposed algorithm.

Algorithm 1 Automatic registration of scans by maximization of mutual information (MI)

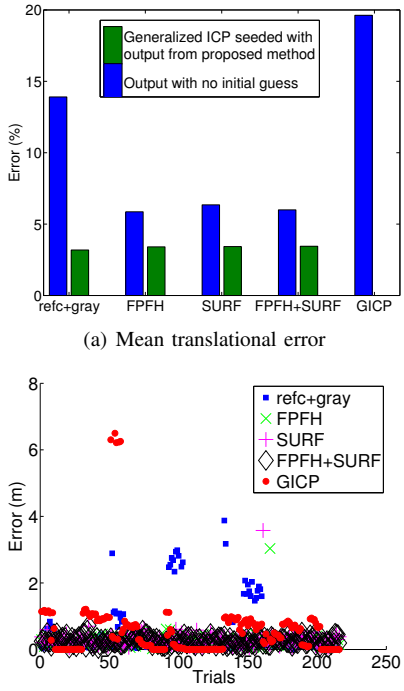
1: **Input:** Co-registered camera and lidar scans $\mathbf{P}$ and $\mathbf{Q}$. Initial guess of the rigid-body transformation $\mathbf{\Theta}_0$.
2: **Output:** Estimated registration parameter $\{\mathbf{\Theta}\}$.
3: Extract generalized feature vectors from scans $\mathbf{P}$ and $\mathbf{Q}$.
4: Quantize the feature vectors using the pre-computed *codebook*.
5: **while** convergence of Nelder-Mead simplex optimization **do**
6: Calculate correspondence of codewords for the current transformation $\mathbf{\Theta}_k$.
7: Calculate the marginal and joint histogram of the corresponding codewords.
8: Calculate shrinkage coefficient λ (7).
9: Calculate James-Stein estimator of the marginal and joint distributions (6).
10: Calculate the MI: $\mathrm{MI}(X, Y; \mathbf{\Theta}_k)$.
11: Update $\mathbf{\Theta}_k \rightarrow \mathbf{\Theta}_{k+1}$.
12: **end while**

A. *Effect of using data from both modalities (camera/lidar)*

In this experiment we demonstrate the effect of choice of features on the robustness of the algorithm. We show that incorporating features from both modalities (camera/lidar) into the registration process improves the performance. We tested our algorithm for the following features:

1) *Reflectivity and Grayscale (refc+gray):* We used approximately 20,000 uniformly sampled points from the textured scan. The reflectivity obtained from the lidar and the corresponding grayscale intensity obtained from the camera are used as a two dimensional feature descriptor.
2) *3D only (FPFH):* Keypoints were detected using the Harris (3D) keypoint detection algorithm available in point cloud library (PCL) [27]. The number of keypoints extracted from a point cloud were between 500–1000.
3) *Image only (SURF):* We used OpenCV's implementation of SURF to extract image keypoints. We assigned the corresponding SURF descriptor to all 3D points that projected within 1-pixel of these keypoints. Only a fraction of the 3D points were assigned these SURF features ($\sim$500–1000).
4) *3D and Image combined (FPFH+SURF):* For all the 3D points that are associated to a SURF descriptor, we calculate the FPFH and append it to the existing SURF descriptor.

In this experiment we randomly selected 200 scan-pairs from the *downtown* dataset spaced approximately 1–5 m apart. We aligned these scan-pairs using the proposed algorithm without any initial guess (i.e., initial guess was fixed at $[0, 0, 0, 0, 0, 0]^\top$). In Fig. 8 we have plotted the translation error in the output of the proposed algorithm for different kinds of features used. We also compared the output of



(a) Mean translational error

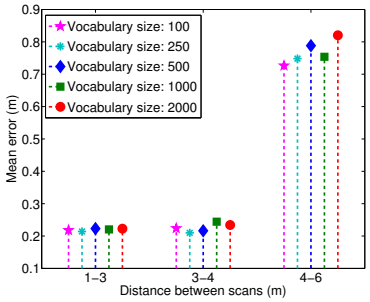
(b) Translational error for each trial

Fig. 8. (a) The blue bars depict the mean registration error starting from no initial guess. The error is calculated as the percentage of the distance between the scans that are aligned (i.e., $error = \frac{\|\mathbf{t}-\hat{\mathbf{t}}\|}{\|\mathbf{t}\|} \times 100$, where $\mathbf{t}$ = true translation vector; $\hat{\mathbf{t}}$ = estimated translation vector; $\|\cdot\|$ = euclidean norm). The green bars represent the mean error for the same set of scans aligned using GICP seeded with the output obtained from the proposed algorithm. The GICP algorithm alone does not converge in the absence of a good initial guess (far right error bar). (b) Here we have plotted the translation error ($\|\mathbf{t}-\hat{\mathbf{t}}\|$) for each trial. The proposed algorithm works well in all trials when we use high-dimensional features. In the case of simple features (refc+gray), the algorithm often gets trapped in a local minima similar to the GICP algorithm (see red circles and blue squares).

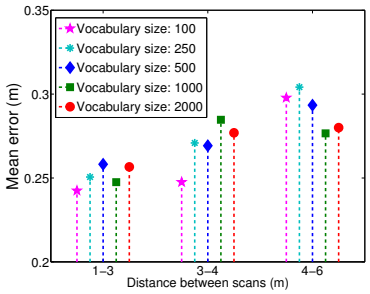
the proposed algorithm with the GICP algorithm that uses the 3D point cloud alone. We found that for a poor initial guess, the GICP algorithm fails to converge whereas the proposed algorithm gives better convergence. As shown in Fig. 8(a) the average error is reduced when we use high-dimensional features instead of simple surface reflectivity values. If we look at the error in each trial (Fig. 8(b)), then we see that the algorithm converges (close to the optimum) in all trials when high-dimensional features are used. However, for simple features (refc+gray), the algorithm is often trapped in a local minima similar to the GICP algorithm (see red circle and blue squares in Fig. 8(b)). The average error in the proposed algorithm can be further reduced by passing its output as an initial guess to the GICP algorithm (the green bars in Fig. 8(a)). Thus, the proposed method provides a principled way to incorporate any kind of features into the registration process that helps in reducing the registration error.

B. *Effect of vocabulary size*

In this experiment we analyze the effect of vocabulary size (i.e., the quantization levels of the codebook) on the proposed algorithm. Since we are not trying to do any recognition, we

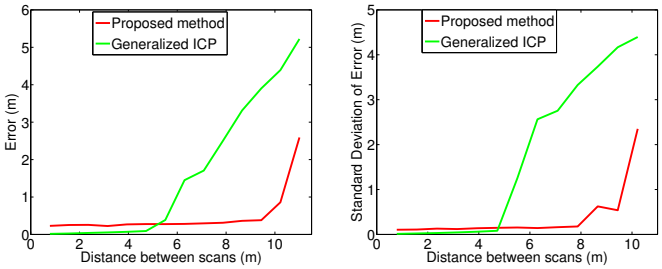


(a) Features used: refc+gray

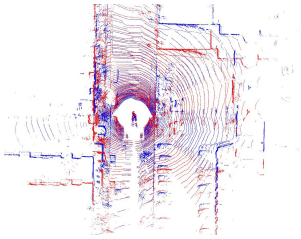


(b) Features used: FPFH+SURF

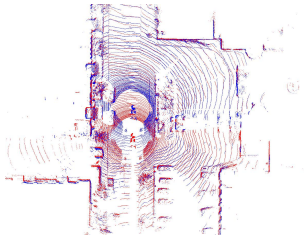
Fig. 9. Mean error in translation is plotted as a function of distance between the scans for different vocabulary sizes of two different features: (a) refc+gray and (b) FPFH+SURF.



(a) Error comparison between GICP and proposed method



(b) Registration result for GICP



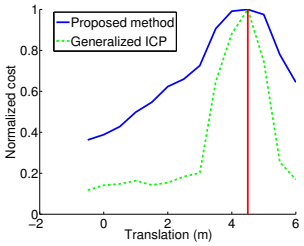
(c) Registration result for MI

Fig. 10. Error comparison between GICP and proposed method with (FPFH+SURF) features. (a) Graph showing the error and standard deviation in translation as the distance between scans **P** and **Q** is increased. (b)–(c) Top view of the 3D scans aligned with the output of GICP and proposed method for two scans that are approximately 6 m apart. Note that the GICP algorithm fails to align the two scans after approximately 4 m whereas the proposed method shows better convergence property and aligns scans that are almost 10 m apart.

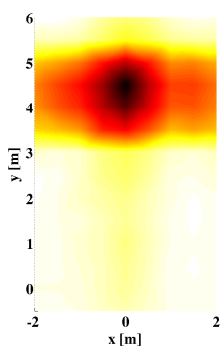
do not need very fine quantization (i.e., large codebook size), and can use a coarse codebook. Moreover, the computation time of our algorithm increases with the size of the codebook. Therefore, we would like to keep the size of the codebook as small as possible. With this experiment we try to identify the optimum size of the codebook for a particular choice of features. We learned the codebook of different sizes $(100, 250, \cdots, 1000)$ for each particular feature set (e.g., refc+gray, FPFH+SURF). We randomly selected 150 scan-pairs (1–3 m, 3–4 m and 4–6 m apart) from the *downtown* dataset. We aligned these scan-pairs using the proposed algorithm (no initial guess) with different codebooks to quantize the features. In Fig. 9 we have plotted the mean translation error for the different codebook sizes. As shown in Fig. 9 the average error increases with the distance between the scans. Although, for simple features (refc+gray), plotted on top panel of Fig. 9, the increase in error is much more than high dimensional features. The effect of vocabulary size as seen in Fig. 9 is dependent upon both features used to create the codebook as well as the distance between the scans under consideration. For example, we found that the optimum value of codebook size for the combined 3D and image features (i.e., FPFH+SURF) is 100 when the distance between the scans is less than 4m, but for larger distances between the scans a finer codebook (vocabulary size = 1000) gives better results. Since the computation complexity of our algorithm is directly proportional to the codebook size, we use smaller codebook sizes as the gain in accuracy is not very large.

C. Comparison with generalized ICP

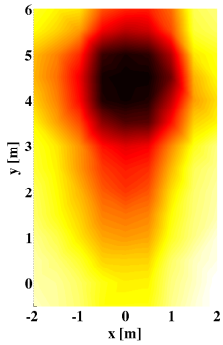
In this experiment we show that the proposed MI-based cost function has a wider basin of convergence as opposed to the state-of-the-art GICP algorithm [11]. Here, we selected a series of 15 consecutive scans from the *downtown* dataset. The average distance between the consecutive scans is approximately 0.5 m–1.0 m. In this experiment we fixed the first scan to be the reference scan and then tried to align the remaining scans (2–15) with the base scan using *(i)* GICP, and *(ii)* our proposed method. The average error in translational motion between the base scan and the remaining scans obtained from these algorithms is plotted in Fig. 10, computed over 90 trials. We found the plotted error trend to be typical across all of our experiments—in general the GICP algorithm alone would fail after approximately 4 m of displacement when not fed an initial guess. The reason for this becomes more clear by analyzing the cost function of the two algorithms. In Fig. 11 we have plotted the cost function of the proposed algorithm and the GICP algorithm for two scans that are 4.5 m apart. Clearly, the proposed method has a wider basin of attraction in both x and y direction. The GICP based cost function (plotted in Fig. 11(b)) has a narrow basin of attraction in the y direction but shows better convergence along the x direction. This is mainly due to the nature of the GICP cost which allows sliding along planar surfaces. The ground plane and the planar structures on both sides of the road (Fig. 11(d)) does not constrain the translation along the y direction but it constrains the motion in x direction. Unlike GICP cost the proposed method does not suffers from planar structures and provides a wider basin of attraction in all directions, thereby converging to the correct solution even if the initial guess is extremely poor. Whereas the GICP cost function shows better convexity near



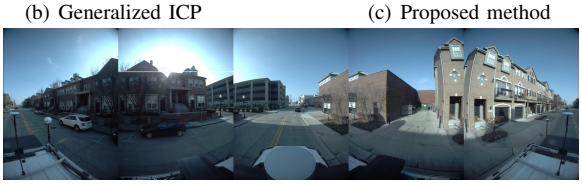
(a) Cost function comparison 1D



(b) Generalized ICP



(c) Proposed method



(d) Image corresponding to the scans for which the cost functions are evaluated above.

Fig. 11. In (a) we have plotted cost as a function of translation along y direction (i.e. the direction of motion of vehicle). In (b) and (c) we have plotted the cost funtion for the same scans by varying both x and y parameters of the rigid body transformation, while keeping the remaining parameters to be fixed to the true value. The proposed method (c) has a wider basin of attraction in both x and y direction, whereas the GICP based cost function (b) has a narrow basin of attraction in the y direction. The basin of attraction in the x direction (b) is better in this case mainly due to the vertical buildings present on both sides (d).

the global maxima, but has a poor basin of convergence. This means the GICP algorithm will converge faster if the initial guess is close to the global maxima but will fail to converge otherwise.

IV. Conclusion

This paper reported on a MI-based scan registration algorithm that allows for the principled fusion of camera and lidar modality information within a single optimization framework. The widespread input flexibility of this algorithm was demonstrated through the use of several different feature sets ranging from very simple (reflectivity + grayscale) to advanced (FPFH+SURF). The proposed algorithm demonstrated good convergence performance and a wider capture basin than state-of-the-art GICP, when implemented with high-dimensional features. GICP, however, shows a more peaked cost function and is able to obtain a lower final registration error when given a good initial guess. Future work will examine how to improve the MI-based cost function so that it shows a sharper response near the optima and thus yields final registration errors comparable to GICP.

References

[1] G. Pandey, J. McBride, S. Savarese, and R. Eustice, "Extrinsic calibration of a 3d laser scanner and an omnidirectional camera," in *IFAC Symp. Intell. Autonomous Vehicles*, 2010.

[2] Q. Zhang, "Extrinsic calibration of a camera and laser range finder," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Sendai, Japan, 2004, pp. 2301–2306.

[3] R. Unnikrishnan and M. Hebert, "Fast extrinsic calibration of a laser rangefinder to a camera," Robotics Institute, Carnegie Mellon University, CMU-RI-TR-05-09, Tech. Rep., July 2005.

[4] P. J. Besl and N. D. McKay, "A method for registration of 3D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, 1992.

[5] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," *Image Vis. Comput.*, vol. 10, no. 3, pp. 145–155, 1992.

[6] M. Alshawa, "ICL: Iterative closest line a novel point cloud registration algorithm based on linear features," *Ekscentar*, no. 10, pp. 53–59, Dec. 2007.

[7] S. Rusinkiewicz and M. Levoy, "Efficient variants of the ICP algorithm," in *Proc. 3-D Digital Imaging and Modeling*, 2001, pp. 145–152.

[8] P. Biber, S. Fleck, and W. Strasser, "A probabilistic framework for robust and accurate matching of point clouds," *Pattern Recognition*, pp. 480–487, 2004.

[9] D. Hahnel and W. Burgard, "Probabilistic matching for 3D scan registration," in *Proc. VDI Conf. Robotik*, 2002.

[10] P. Biber, "The normal distribution transform: A new approach to laser scan matching," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, vol. 3, 2003, pp. 2743–2748.

[11] A. V. Segal, D. Haehnel, and S. Thrun, "Generalized-ICP," in *Proc. Robot.: Sci. & Syst. Conf.*, 2009.

[12] A. Johnson and S. B. Kang, "Registration and integration of textured 3D data," in *Image and Vision Computing*, 1996, pp. 234–241.

[13] D. Akca, "Matching of 3D surfaces and their intensities," *ISPRS J. Photogrammetry and Remote Sensing*, vol. 62, pp. 112–121, 2007.

[14] G. Pandey, J. R. McBride, S. Savarese, and R. M. Eustice, "Visually bootstrapped generalized ICP," in *Proc. IEEE Int. Conf. Robot. and Automation*, Shanghai, China, May 2011, pp. 2660–2667.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[16] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proc. European Conf. Comput. Vis.*, 2006, pp. 404–417.

[17] P. Viola and W. Wells, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, pp. 137–154, 1997.

[18] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, pp. 187–198, 1997.

[19] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual information based registration of medical images: A survey," *IEEE Trans. Med. Imag.*, vol. 22, no. 8, pp. 986–1004, 2003.

[20] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *Proc. IEEE Int. Conf. Robot. and Automation*, Kobe, Japan, May 2009, pp. 3212–3217.

[21] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, pp. 1265–1278, Aug. 2005.

[22] A. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3D scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 5, pp. 433 –449, May 1999.

[23] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, vol. 2, 2006, pp. 2161–2168.

[24] J. Hausser and K. Strimmer, "Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks," *J. Mach. Learning Res.*, vol. 10, no. Dec., pp. 1469–1484, 2009.

[25] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, pp. 308–313, 1965.

[26] G. Pandey, J. R. McBride, and R. M. Eustice, "Ford campus vision and lidar data set," *Int. J. Robot. Res.*, vol. 30, no. 13, pp. 1543–1552, Nov. 2011.

[27] R. B. Rusu and S. Cousins, "3D is here: Point cloud library (PCL)," in *Proc. IEEE Int. Conf. Robot. and Automation*, May 2011, pp. 1–4.