Toward Mutual Information Based Place Recognition

Gaurav Pandey[1], James R. McBride[2], Silvio Savarese[3] and Ryan M. Eustice[4]

***Abstract*— This paper reports on a novel mutual information (MI) based algorithm for robust place recognition. The proposed method provides a principled framework for fusing the complementary information obtained from 3D lidar and camera imagery for recognizing places within an *a priori* map of a dynamic environment. The visual appearance of the locations in the map can be significantly different due to changing weather, lighting conditions and dynamical objects present in the environment. Various 3D/2D features are extracted from the textured point clouds (scans) and each scan is represented as a collection of these features. For two scans acquired from the same location, the high value of MI between the features present in the scans indicates that the scans are captured from the same location. We use a non-parametric entropy estimator to estimate the true MI from the sparse marginal and joint histograms of the features extracted from the scans. Experimental results using seasonal datasets collected over several years are used to validate the robustness of the proposed algorithm.**

I. Introduction

Today, robots are required to operate in an environment for days, months or even years. One important task that any robot needs to perform in these long-term environments is to recognize places it has visited before. This place recognition capability has a wide range of applications in autonomous navigation including global localization and loop-closure detection for simultaneous localization and mapping (SLAM) [1]–[3]. The task of place recognition in a dynamic environment becomes extremely challenging as a single location appears different over time. The drastic changes in environmental appearance due to changing seasons (summer, fall, winter, etc.), lighting conditions, and dynamical objects make the task of place recognition very challenging (Fig. 1).

Most place recognition literature has focused on obtaining correct loop-closures for SLAM. In these situations, the robot creates a map of an *a priori* unknown environment while simultaneously localizing itself in this map. Therefore, the robot has to recognize a place that has been recently visited or added to the map. The time difference between the two instances is usually small and hence the change in appearance of the environment is not too large (apart from change in viewpoint). Vision-based algorithms based on Bag-of-Words techniques [4], [5] have been successfully used for robust place recognition in scenarios like this. Cummins et al. [6] presented a probabilistic framework, Fast Appearance-Based Mapping (FAB-MAP), that is robust to perceptual aliasing for appearance-based place recognition over maps as big as 1000 km long. Pronobis et al. [7] described a fully supervised method for place recognition that is robust to different illumination conditions in indoor scenes. Sunderhauf and Protzel [8] proposed a simple appearance-based place recognition system based on Binary Robust Independent Elementary Feature (BRIEF) descriptors and showed that its performance is comparable to FAB-MAP for large-scale SLAM problems.

Recently, the problem of long-term navigation in a changing environment has received significant attention in the mobile robotics community. The ability to recognize places across seasons, with significant appearance changes (e.g., Fig. 1) is very important for long-term autonomy. Glover

Fig. 1. Sample imagery extracted from three different datasets captured in December 2009, October 2010 and February 2011; each row corresponds to the same place. The datasets exhibit significant visual changes due to different weather conditions, lighting and dynamical objects.

*This work was supported by Ford Motor Company via the Ford-UM Alliance under award N015392.

[1]G. Pandey is with the Department of Electrical Engineering & Computer Science, University of Michigan, Ann Arbor, MI 48109, USA pgaurav@umich.edu.

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

[3]S. Savarese is with the Department of Computer Science, Stanford University, Stanford, CA 94305, USA ssilvio@stanford.edu.

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

et al. [9] presented a combination of FAB-MAP [6] and the biologically inspired RatSLAM [10] approach, and showed that it is robust to illumination and structural changes in outdoor environments. Milford et al. [11] proposed to match sequences of images instead of a single image and showed good precision in recognizing places across different seasons (summer-rain). Churchill and Newman [12] introduced the concept of plastic maps (i.e., a composite representation constructed from multiple overlapping experiences). As the robot repeatedly travels through the same environment under different conditions, it accumulates distinct visual experiences that represent the scene variation. They have shown good results on a road vehicle operating over a three month period at different times of day, in different weather, and different lighting conditions. Neubert et al. [13] proposed a novel idea of appearance change prediction. They learn the change in the visual appearance of the environment over time and then use this learned knowledge to predict the appearance of any place under different environmental conditions.

The methods mentioned so far are purely vision-based and use camera as the primary sensing modality. However, robots are often equipped with various perception sensors besides camera like lidar, radar, etc. Although these sensors provide useful complementary information to the camera data, they are seldom used for place recognition. There have been some attempts to increase the robustness of place recognition in SLAM systems by fusing the multi-modal data at the landmark level [14]. Paul and Newman developed a more robust FAB-MAP 3D algorithm for large-scale SLAM systems [15] by extending the appearance-only FAB-MAP algorithm to incorporate spatial information of the visual features obtained from laser scanners.

Most of the aforementioned methods either use the image 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—as 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. [16] and Maes et al. [17]. Since then, researchers (especially in medical imaging) have widely used the MI framework to focus on specific registration problems in various clinical applications [18]. 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 place recognition using co-registered 3D lidar and camera imagery (Fig. 2). Our method provides a robust framework for incorporating complementary information obtained from these modalities into the recognition process.

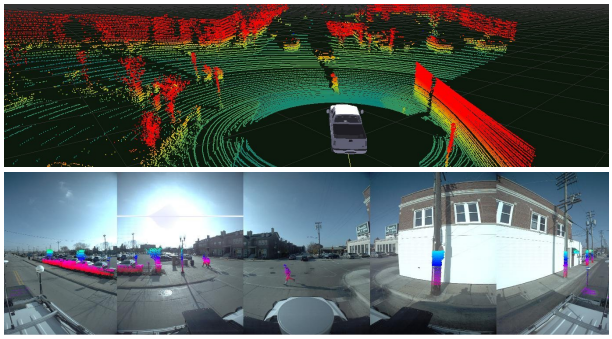


Fig. 2. The top panel is a perspective view of the Velodyne 3D lidar range data, color-coded by height above the ground plane. The bottom panel shows the above ground plane range data projected into the corresponding image from the Ladybug3 camera. Several recognizable objects are present in the scene (e.g., people, stop signs, lamp posts, trees). Only nearby objects are projected for visual clarity.

The remainder of this paper proceeds as follows: In Section II we describe the proposed method of automatic place recognition. In Section III we present results showing the robustness of the proposed method and present a comparison against a standard Bag-of-Words approach. Finally, in Section IV we summarize our findings.

II. Methodology

In our work, we have used data from a 3D laser scanner and an omnidirectional camera system mounted on a mobile robotic platform specifically designed for long-term autonomous navigation in a dynamic environment [19]. The robot travels through the environment in different seasons (summer, fall, winter) and captures time synchronized lidar and camera data. We assume that the intrinsic and extrinsic calibration parameters for these sensors are either known or estimated beforehand (e.g., using algorithms such as [20], [21]).

The calibration of sensors allows us to project 3D points from lidar onto the corresponding camera image (and vice versa), as shown in Fig. 2. This co-registration allows us to associate features extracted from the camera image (grayscale value, scale invariant feature transform (SIFT) [22], speeded up robust features (SURF) [23], etc.) to the corresponding 3D lidar point that projects onto that pixel location. The features extracted from the 3D point cloud (reflectivity, normals, etc.) and camera image are fused together (discussed later in section II-B) and every scan is represented as a collection of these features. Thus, for any two scans corresponding to the same physical location, the joint distribution of these features should show maximum correlation. Here, we use MI as a measure of this correlation and a simple thresholding scheme to localize the scans within a prior map. An overview of the proposed method is given in Fig. 3.

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

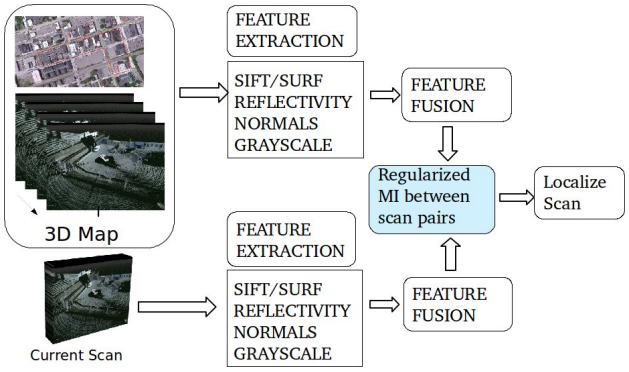


Fig. 3. Overview of the proposed algorithm.

which demonstrate a measure of statistical dependence of the random variables in consideration. One such form of MI is defined in terms of the 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. Sensor Data Fusion

In this section we describe two novel techniques to fuse the various features extracted from the co-registered lidar/camera data. A mobile robot equipped with 3D lidar and camera drives through the environment and captures time-aligned lidar and camera data. We extract both simple features (e.g., reflectivity, grayscale) and high dimensional features (e.g., SIFT, SURF) from this data. It is important to note that simple features, like the reflectivity of the 3D points obtained from the lidar, or the color of the pixel obtained from the camera, are discrete signals generated by sampling the same physical scene but in a different manner. Since the underlying structure generating these signals is the same, they are statistically dependent upon each other and can be fused together at the *signal level*; whereas the high dimensional features (such as SIFT/SURF) from imagery are generally independent from the reflectivity of the lidar point and are therefore fused at the *information level*.

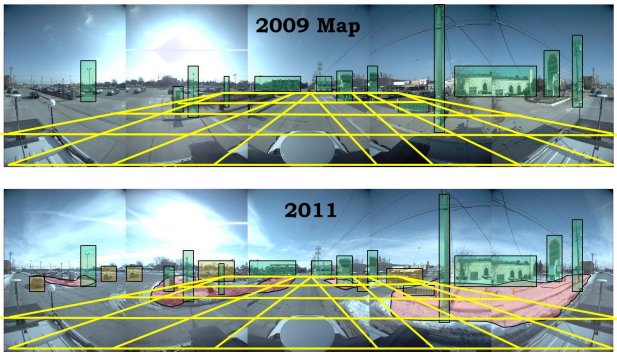


Fig. 4. The top panel shows the omnidirectional image of a location captured in fall 2009. The bottom panel shows the omnidirectional image of the same location in winter 2011. The significant change in the scene is clearly visible from the two images, for example, snow on the ground (marked in red), dynamic objects (marked in orange), lighting conditions, etc. Such drastic changes make registration of the 2009 and 2011 datasets a challenging problem. However, there are also common objects (marked in green) that have stationary statistics and can be used for registration of sensor data. The 3D space around the sensor is divided into voxels of equal size. Here we have illustrated the voxelization process in the image via yellow check pattern, however actual voxels are 3-dimensional. The voxelization allows us to use stationary statistics within each voxel for registration.

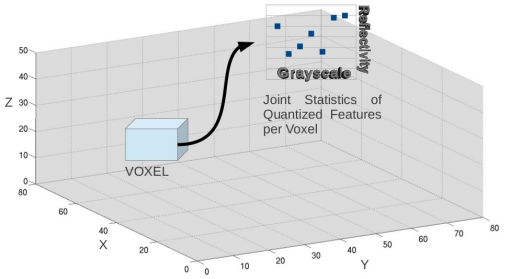


Fig. 5. Sensor data fusion at the *signal level*. The joint statistics of the quantized features present in each voxel constitute the marginal distribution of the features present in the scan.

1) Sensor Data Fusion at the Signal Level: In this section we describe a novel method of fusing lidar/camera data at the signal level. The reflectivity from lidar and grayscale intensity from the camera are measurements generated by the same underlying physical scene. These two modalities are therefore highly correlated (i.e., a highly reflective point in lidar data will typically have a high grayscale value for the corresponding pixel). In order to fuse such highly correlated features we divide the 3D scan into voxels (Fig. 4) of fixed dimension and calculate the joint-statistics of these simple features extracted from lidar/camera data in each voxel in the form of a multi-dimensional histogram (Fig. 5). This multi-dimensional histogram represents the marginal distribution of the fused features present in the scan, which is later used for estimation of MI. The voxelization of scene allows us to use stationary statistics within each voxel for robust registration. When we consider the statistics of features across two scans captured at two different time (fall 2009 and winter 2011) the local stationary statistics shows higher correlation as compared to the global correlation of the entire scene.

2) Sensor Data Fusion at the Information Level: In the previous section we described a method of fusing sensor data that exhibit some correlation (e.g., reflectivity from lidar and grayscale intensity from camera). However, there are also high-dimensional features (e.g., SIFT, SURF) extracted from the camera data that do not necessarily show any correlation with the reflectivity from the lidar data. This is mainly because these high-dimensional features are computed from the grayscale values of the local neighborhood of the pixel, which is generally quite different from the grayscale value of the pixel itself. Therefore, although the grayscale values and the corresponding lidar reflectivity values show high correlation, the corresponding high-dimensional features are generally not correlated with the lidar intensity value. For example, consider a textured surface that has a distinct SIFT feature (128 dimensional vector) at a given pixel (mainly because of the gradient around the pixel), the reflectivity of the 3D point projected onto that pixel is a single value between [0, 255], which does not contain the neighborhood information and is not necessarily related to the image feature. Therefore, we consider such features to be statistically independent of each other and hence do not fuse them at the signal level; however, they contain useful information necessary for place recognition. Therefore, we propose to fuse these features at the information level by simply computing the total MI between any two scans as the sum of mutual information of each of these independent features,

$$\mathrm{TMI}(X,Y) = \sum_i \mathrm{MI}(F_i^X, F_i^Y), \tag{5}$$

where $\mathrm{TMI}(X,Y)$ is the total MI between the scans X and Y, and F_i^X and F_i^Y are various features (fused or independent) extracted from the scan data. It should be noted that if we compute some high-dimensional SIFT/SURF like features from the local neighborhood of the lidar point then they are more likely to be correlated with the corresponding image features and can be fused at *signal* level (after quantization) as shown in previous section (§II-B.1).

C. Mapping and Place Recognition

We first create a map of the environment from the sensor data. The map consists of equally-spaced scans with known location in a global reference frame. Each scan in the map is a collection of quantized features extracted from the sensor data. Simple features, like the reflectivity from lidar and the grayscale intensity values from camera data, are integer values and, therefore, easy to quantize between a given range (generally [0–255] for 8-bit sensors). However, for high-dimensional features (SIFT, SURF, etc.) we first create a dictionary of *codewords* representing the quantization of these features extracted from the scans. We extract N such features (training samples) from a set of scans called the training dataset (Fig. 6). We use a hierarchical k-means clustering [5] 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.

(a) Sample images from the training dataset (*Ford Campus*)

(b) Sample images from the testing dataset (*Downtown*)

Fig. 6. The codebook is 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 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.

We consider the collection of these codewords present in a scan (extracted from the map) as the random variable X. In a given map we have N such scans representing a unique place in the map. The goal of place recognition is to identify the correct location of the robot when it revisits a place in an *a priori* map. Here we assume that the map is created once and the robot revisits some place in the map after a significant amount of elapsed time. We consider the collection of codewords extracted from this scan (which we will refer to as the query scan) as the random variable 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. Let $\mathbf{Q}$ be the query scan and $\mathbf{P}$ be one of the scans in the map. 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 $\mathbf{P}$ and $\mathbf{Q}$, 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 $\mathbf{P}$ onto the reference frame of scan $\mathbf{Q}$ is given by:

$$\hat{\mathbf{q}}_i = \mathrm{R}\mathbf{p}_i + \mathbf{t}. \tag{6}$$

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. 7). A codeword c_i^p in scan $\mathbf{P}$ is first transformed to the reference frame of $\mathbf{Q}$. All the codewords in scan $\mathbf{Q}$ that are within a sphere of radius r around c_i^p are considered as potential correspondences. The codeword c_i^q

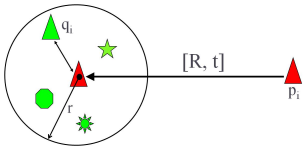


Fig. 7. Illustration of the nearest neighbor search algorithm used to establish codeword correspondence; each shape above represents a different codeword—green colorings belong to scan **Q** and red to scan **P**. All the codewords in scan **Q** that are within a sphere of radius r around c_i^p are considered as potential correspondences. The codeword c_i^q that gives the maximum similarity score with c_i^p is chosen as the correspondence.

Fig. 8. The left panel shows a sample codeword (on the lamp-post) extracted from the query image. The right panel shows the epipolar line (green) for the same codeword in the corresponding image from the map. The potential correspondences are marked in blue and the correct correspondence computed based on codeword similarity is marked in red.

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 the method described above when the 3D location of the codewords is known. However, for certain image features (e.g., SIFT, SURF), their 3D location are often not known due to the sparseness of data obtained from the lidar or due to limited overlap between the field of view of the two sensors. In that case we use the epipolar constraint [24] to establish the correspondence between image features. If $C^P = \{c_i^p; i = 1, \cdots, n\}$ and $C^Q = \{c_i^q; i = 1, \cdots, m\}$ are the set of codewords present in images I^P and I^Q corresponding to scans **P** and **Q**, and $[\mathrm{R}, \mathbf{t}]$ are the rotational and translation parameters between the two cameras, then the two corresponding codewords are related by the epipolar constraint:

$$\tilde{\mathbf{p}}_i^\top \mathrm{F} \tilde{\mathbf{q}}_i = 0, \tag{7}$$

where $\tilde{\mathbf{p}}_i$ and $\tilde{\mathbf{q}}_i$ are the homogeneous pixel coordinates of the codewords c_i^p and c_i^q, respectively. F is the *fundamental* matrix that maps the codeword in image I^P to the corresponding epipolar line in the image I^Q (Fig. 8). Therefore, all points within a certain distance of the epipolar line are considered a potential correspondence and the codeword c_i^q that gives the maximum similarity score with c_i^p is taken to be the true correspondence.

We use this correspondence to create the joint histogram of codewords for the given transformation $[\mathrm{R}, \mathbf{t}]$. The maximum likelihood estimate (MLE) of the marginal and joint probabilities of the random variables X and Y can be obtained from the normalized marginal and joint histograms of these codewords. It is important to note that the number of different codewords present in any scan is generally (especially for high-dimensional features) only a fraction of the size of the codebook. For instance, since the maximum range of the Velodyne laser scanner is $100\ \mathrm{m}$ and the vertical field of view (FOV) of the sensor is $20°$, the dimensions of the viewing cube around the sensor becomes $[200\ \mathrm{m} \times 200\ \mathrm{m} \times 50\ \mathrm{m}]$. If we use voxels of size $1\ \mathrm{m}$ and consider the lidar reflectivity and grayscale intensity values quantized between [0, 255], the size of the histogram that needs to be created becomes extremely large ($200 \times 200 \times 50 \times 256 \times 256 = 131,072,000,000$ bins). The total number of points (n) extracted from a single scan is typically much less than the dimensions of this huge joint histogram. 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 use the Chao-Shen estimator for regularized entropy estimation [25]. This technique has been successfully used in estimating entropy of gene data in an under-sampled regime with missing species in the observed data. In this approach the entropy of the random variable (with few observations, $n \ll K \times K$) is estimated by applying the Horvitz-Thompson estimator [26] in combination with the Good-Turing correction [27] of the MLE. The Good-Turing-corrected probability estimates are given by:

$$X_k^{GT} = (1 - \frac{m_1}{n}) X_k^{ML}, \tag{8}$$

where m_1 is the number of bins with single observation (i.e. $x_k = 1$ and X_k^{ML} is the maximum likelihood (ML) estimate). Combining this with the Horvitz-Thompson estimator, the required entropy is:

$$H^{CS} = -\sum_{k=1}^{n} \frac{X_k^{GT} \log(X_k^{GT})}{(1 - (1 - X_k^{GT})^n)}. \tag{9}$$

Once we have a good estimate of the joint and marginal entropies, we can write the total MI of the features present in the two scans as a function of the rigid-body transformation between the scan pair:

$$\mathrm{TMI}(X, Y; \mathbf{\Theta}) = \sum_i \mathrm{MI}(F_i^X, F_i^Y; \mathbf{\Theta}), \tag{10}$$

where $\mathbf{\Theta} = [x, y, z, \phi, \theta, \psi]^\top$ is the six degree of freedom (DOF) parametrization of the rigid-body transformation $[\mathrm{R}, \mathbf{t}]$. This rigid-body transformation is unknown in the absence of any inertial measurement unit (IMU) or global positioning system (GPS) device. Here we assume that the robot motion is mostly planar, so for every query scan the corresponding scan in the map should be acquired from the same location within a few meters in the x-y plane. Therefore, we perform a linear search over all the scans present in the map dataset with $\mathbf{\Theta} = [0, 0, 0, 0, 0, 0]^\top$ as the transformation parameter. Since we assume planar motion of the vehicle, we also search over certain discrete values of the heading angle (ψ) of the transformation parameters. During this linear search if the TMI is greater than a certain threshold, then we optimize the total MI over the full 6-DOF

rigid-body transformation:

$$\hat{\mathbf{\Theta}} = \underset{\mathbf{\Theta}}{\operatorname{argmax}} \sum_j \mathrm{MI}(F_j^X, F_j^Y; \mathbf{\Theta}), \tag{11}$$

thereby obtaining the exact location of the query scan in the map. We use the simplex method proposed by Nelder and Mead [28] to estimate the optimum value of the registration parameter, $\mathbf{\Theta}$, that maximizes the cost function given in (11). This process is repeated for all the scans in the map and the scan that gives the maximum value of total mutual information with respect to the query scan corresponds to the desired location. The computational complexity of the proposed algorithm depends upon the size of the map that is being searched. Since we do not assume any prior knowledge of the location of the query scan from odometry or any other source, the linear search gets computationally very expensive. The main emphasis of this work though is to show the robustness of a framework that allows to use multi-modal data for recognizing places under significant changes in the appearance of the environment due to changes in weather, lighting, dynamical objects, etc. The complete place recognition method is summarized in Algorithm 1.

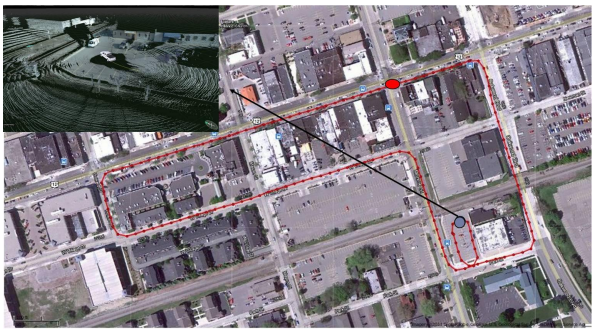
Algorithm 1 Mutual Information based Place Recognition

Input: Co-registered camera and lidar scans $[\mathbf{P}]_i^N$ constituting the map and query scan $\mathbf{Q}$.
Output: Scan index from the map that is closest to query scan $\{\mathrm{INDEX}\}$ and its estimated registration parameter $\{\hat{\mathbf{\Theta}}\}$.
Extract generalized feature vectors from query scan $\{F^Y\}_j^M$.
Quantize and fuse features.
Let $\mathrm{MAX} = \mathrm{THRESHOLD}$, $\mathrm{INDEX} = 0$;
while i = 1 to N **do**
  Get the quantized feature vectors from map $\{F^X\}_j^M \leftarrow \mathbf{P}_i$.
  for $\psi = 0 : 60^\circ : 360^\circ$ **do**
    $\mathbf{\Theta} \leftarrow [0, 0, 0, 0, 0, \psi]^\top$;
    Calculate the total MI:
    $\mathrm{TMI} = \sum_j \mathrm{MI}(F_j^X, F_j^Y; \mathbf{\Theta})$;
    if $\mathrm{TMI} \geq \mathrm{MAX}$ **then**
      $\hat{\mathbf{\Theta}} = \underset{\mathbf{\Theta}}{\operatorname{argmax}} \sum_j \mathrm{MI}(F_j^X, F_j^Y; \mathbf{\Theta})$
      $\mathrm{MAX} = \sum_j \mathrm{MI}(F_j^X, F_j^Y; \hat{\mathbf{\Theta}})$
      $\mathrm{INDEX} = i$;
    end if
  end for
end while

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 (Fig. 9(a)). We use the pose information available from a high end IMU (Applanix POS-LV 420 INS with Trimble GPS) as the ground-truth to compare the place recognition errors. The dataset used in our experiments are divided into two distinct runs: *(i) Downtown* and *(ii) Ford Campus*, both taken in Dearborn, Michigan [19]. We have several different sets of data recorded at different times of the year from these locations. One such set recorded in December 2009 is also available online. In our experiments

(a) Test vehicle (left). The 3D laser scanner and omnidirectional camera system mounted on the roof of the vehicle (right) as described in [19]



(b) 3D map of a section of downtown Dearborn created from the data collected in December 2009. Each node in the map is comprised of a textured 3D point cloud representing a distinct place in the map.

Fig. 9. The top panel shows the autonomous vehicle platform and the bottom panel shows a section of the map generated from the data collected from the sensors mounted on the vehicle.

we have used the *Downtown* dataset for testing and the *Ford Campus* dataset for learning the codebook. We have used five different runs of the *Downtown* dataset recorded in December 2009, September 2010, October 2010, February 2011 and March 2011 for testing. Each of these runs exhibit significant changes due to weather (e.g., snow on the ground in 2011, no leaves on the trees in December 2009), construction (road blocked, trailers parked) and lighting, etc, thereby making place recognition a challenging task. In our experiments we have used the December 2009 dataset as the prior map (Fig. 9(b)) and used scans from the other four datasets as query scans for place recognition. The map dataset contains about 500 equi-spaced scans (approximately 10 m apart), where each scan represents a unique location. We performed the following experiments to analyze the robustness of the proposed algorithm.

A. Effect of Using Data from Both Camera and 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 performance. We tested our algorithm for the following features:

- *Reflectivity and Normal:* The reflectivity of the point obtained from lidar and the surface normal at the point are used as features. They are assumed to be independent and fused together at the information level as described in §II-B.2
- *Reflectivity, Grayscale and Normal:* The reflectivity and corresponding grayscale value of a 3D point show high

correlation and are fused at the signal level as described in §II-B.1. The combined reflectivity and grayscale feature is then fused with the extracted surface normals at the information level (§II-B.2).

- *SURF:* We use OpenCV's [29] implementation of the SURF feature detector and descriptor. It should be noted that we utilize the 3D location of these SURF features to establish correspondences as described in §II-C. Therefore, it should not be confused with pure vision-based technique since we are accounting for the 3D location of these features coming from the lidar data.
- *Reflectivity, Normal and SURF:* Here we combine the SURF features with the 3D features (reflectivity and normal). Since SURF features are completely independent of the reflectivity or normal of the 3D point, these features are fused at the information level.

Here we created a prior map from the *Downtown* dataset recorded in December 2009, scans from the data recorded in 2010 and 2011 are treated as query scans. The December 2009 data corresponds to a typical *winter* day with no snow on the ground anywhere and trees without any leaves. We used the scans from the data recorded in 2010 and 2011 as query scans. The query scan is aligned with each scan in the map and the one that gives the highest value of total MI is considered as the best match. In Fig. 10(a) and (b) we have plotted the precision-recall (PR) curves for the data collected in September and October 2010, respectively. If the scan from the map that gives the highest MI is within $1\ \mathrm{m}$ of the query scan then it is considered as a true positive. We use the GPS information available from the IMU as ground truth to calculate the correct distance between any two scans. The query dataset here is quite different from the *winter* map dataset not only due to change in dynamical objects (cars parked on the roads/parking lot, etc.) but due to change in weather also. The trees in this dataset are filled with leaves unlike the map dataset. Similarly, in Fig. 10(d) and (e) we have plotted the precision-recall curves for data collected in February and March 2011, respectively. These datasets collected in winter have snow on the ground and hence exhibit significant change in the appearance of the same location as compared to the map dataset. In both cases we observed that the precision of the proposed algorithm increases as we increase the complexity of features (i.e., using high-dimensional SURF features improves the performance of the algorithm). We also observed an increase in performance when we incorporated data from both modalities.

B. Comparison with Bag of Words Method

Here we compare the output of the proposed algorithm with the standard bag-of-words algorithm proposed in [5]. We used the same training dataset for learning the vocabulary for both methods. We observe that the proposed algorithm outperforms the bag-of-words algorithm, which is not surprising since our algorithm takes full advantage of the additional laser modality. In Fig. 10(c) and (f) we have plotted the precision-recall curve for 2010 and 2011 datasets, respectively, for both of the methods. The performance of the proposed algorithm (the best output that uses reflectivity, normals and SURF together) is significantly higher as compared to the bag-of-words method. This is mainly because the bag-of-words algorithm only uses the images and does not exploit the 3D information available from the lidar data.

IV. Conclusion and Future Works

This paper reported on a MI-based place recognition algorithm that allows for the principled fusion of camera and lidar modality information within a single framework. The proposed algorithm showed good results for real data collected from an autonomous vehicle platform, over a period of 3 years at different times of day, under different weather conditions, and with significant lighting and structural changes. The proposed method outperformed the standard image-based technique (bag-of-words) used for place recognition. The ability of the proposed algorithm to recognize places across seasons, with significant appearance changes makes it very suitable for long-term autonomous operation of robots. Future work includes using the proposed algorithm as a front-end for loop-closure detection in a real-time SLAM system. We also intend to compare the proposed algorithm with other state-of-the-art methods for place recognition.

References

[1] K. L. Ho and P. Newman, "Loop closure detection in SLAM by combining visual and spatial appearance," *Robotics and Autonomous Systems*, vol. 54, pp. 740–749, July 2006.

[2] J. Callmer, K. Granstrom, J. Nieto, and F. Ramos, "Tree of words for visual loop closure detection in urban SLAM," in *Proceedings of the Australasian Conference on Robotics and Automation*, Canberra, Australia, Dec. 2008, pp. 102–110.

[3] A. Angeli, S. Doncieux, J. Meyer, and D. Filliat, "Visual topological SLAM and global localization," in *Proc. IEEE Int. Conf. Robot. and Automation*, Kobe, Japan, May 2009, pp. 4300–4305.

[4] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1470–1477.

[5] 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.

[6] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *Int. J. Robot. Res.*, vol. 30, no. 9, pp. 1100–1123, 2011.

[7] A. Pronobis, B. Caputo, P. Jensfelt, and H. I. Christensen, "A discriminative approach to robust visual place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, 2006, pp. 3829–3836.

[8] N. Sunderhauf and P. Protzel, "BRIEF-Gist—Closing the loop by simple means," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, 2011, pp. 1234–1241.

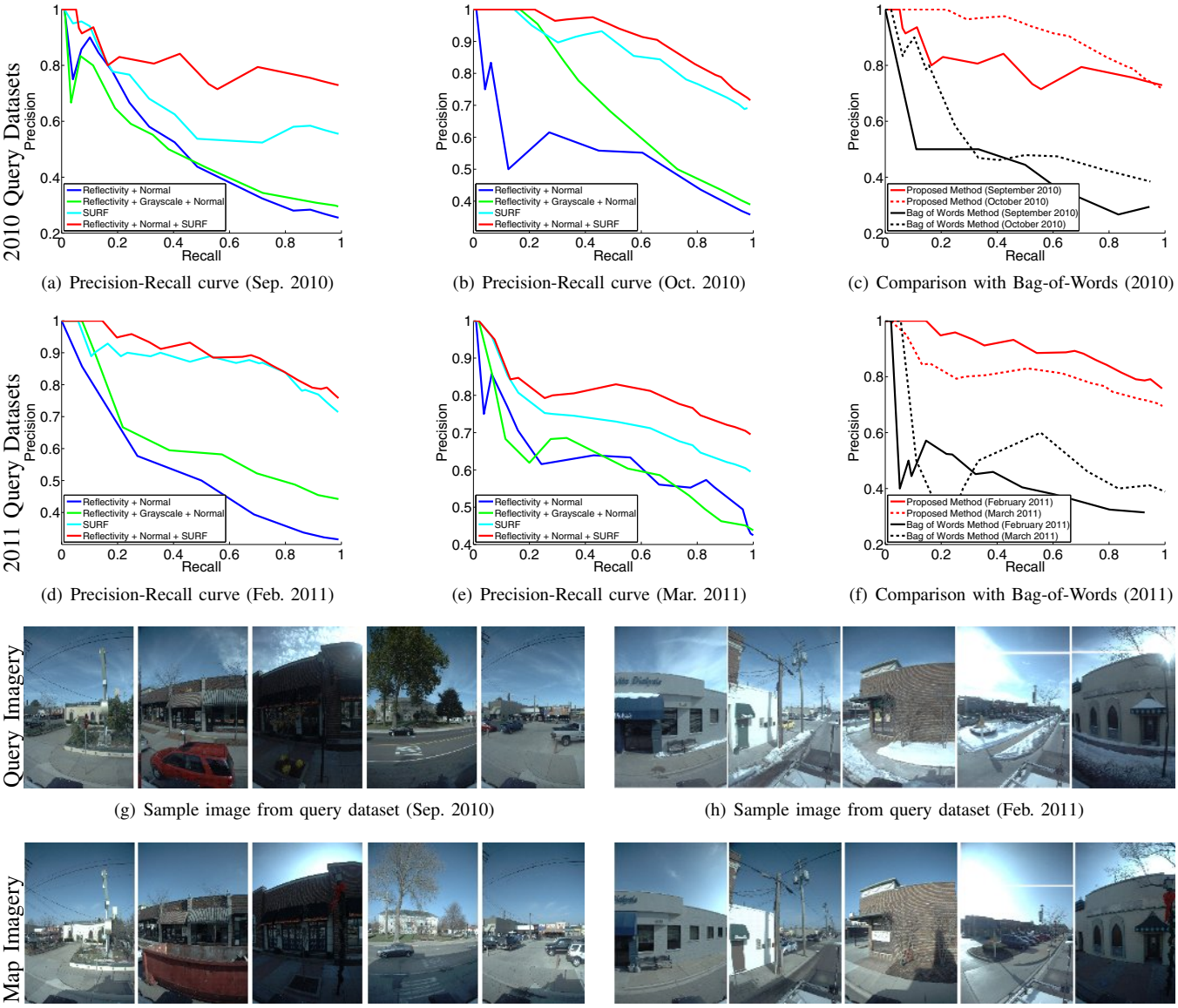
[9] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth, "FAB-MAP + RatSLAM: Appearance-based SLAM for multiple times of day," in *Proc. IEEE Int. Conf. Robot. and Automation*, 2010, pp. 3507–3512.

[10] M. Milford, G. Wyeth, and D. Prasser, "RatSLAM: A hippocampal model for simultaneous localization and mapping." in *Proc. IEEE Int. Conf. Robot. and Automation*, 2004, pp. 403–408.

[11] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proc. IEEE Int. Conf. Robot. and Automation*, 2012, pp. 1643–1649.

[12] W. Churchill and P. Newman, "Practice makes perfect? Managing and leveraging visual experiences for lifelong navigation," in *Proc. IEEE Int. Conf. Robot. and Automation*, 2012, pp. 4525–4532.

[13] P. Neubert, N. Sunderhauf, and P. Protzel, "Appearance change prediction for long-term navigation across seasons," in *Proc. European Conf. Mobile Robots*, Sept. 2013, Accepted, To Appear.



(a) Precision-Recall curve (Sep. 2010)

(b) Precision-Recall curve (Oct. 2010)

(c) Comparison with Bag-of-Words (2010)

(d) Precision-Recall curve (Feb. 2011)

(e) Precision-Recall curve (Mar. 2011)

(f) Comparison with Bag-of-Words (2011)

(g) Sample image from query dataset (Sep. 2010)

(h) Sample image from query dataset (Feb. 2011)

(i) Corresponding correct image retrieved from map dataset (Dec. 2009)

(j) Corresponding correct image retrieved from map dataset (Dec. 2009)

Fig. 10. Precision-Recall curves and sample images from the query (2010, 2011) and map (2009) datasets.

[14] J. A. Castellanos, J. Neira, and J. D. Tardos, "Multisensor fusion for simultaneous localization and map buildiing." *IEEE Trans. Robot. Autom.*, vol. 17, pp. 908–914, 2002.

[15] R. Paul and P. Newman, "FAB-MAP 3D: Topological mapping with spatial and visual appearance," in *Proc. IEEE Int. Conf. Robot. and Automation*, 2010, pp. 2649–2656.

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

[17] 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.

[18] 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.

[19] 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.

[20] G. Pandey, J. R. McBride, S. Savarese, and R. M. Eustice, "Automatic targetless extrinsic calibration of a 3D lidar and camera by maximizing mutual information," in *Proc. AAAI Nat. Conf. Artif. Intell.*, Toronto, Canada, July 2012, pp. 2053–2059.

[21] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.

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

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

[24] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[25] A. Chao and T. J. Shen, "Nonparametric estimation of Shannons index of diversity when there are unseen species in sample," *Environmental and Ecological Statistics*, vol. 10, no. 4, pp. 429–443, 2003.

[26] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe." *J. American Statistical Assoc.*, vol. 47, pp. 663–685, 1952.

[27] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Always good turing: Asymptotically optimal probability estimation," *Science*, vol. 302, pp. 427–431, 2003.

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

[29] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.