2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Learning Pedestrian Activities for Semantic Mapping

B. Qin[1], Z. J. Chong[1], T. Bandyopadhyay[2], M. H. Ang Jr.[1], E. Frazzoli[3], D. Rus[3]

Abstract— **This paper proposes a semantic mapping method based on pedestrian activity in the urban road environment. Pedestrian activity patterns are learned from pedestrian tracks collected by a mobile platform. With the learned knowledge of pedestrian activity, semantic mapping is performed using Bayesian classification techniques. The proposed method is tested in real experiments, and shows promising results in recognizing four activity-related semantic properties of the urban road environment: pedestrian path, entrance/exit, pedestrian crossing and sidewalk.**

I. Introduction

Semantic mapping has become a popular research topic in recent years. By augmenting the traditional metric/topological maps with higher-level semantic knowledge, researchers aim to help robots to really "understand" their environments. A semantic map can not only facilitate human-robot interaction, but also help a robot perform advanced reasoning and planning. In the past few years, various methods have been proposed for semantic mapping. Depending on the sources of semantic information, these methods can be roughly classified into three categories: appearance-based approach, object-based approach, and activity-based approach.

The appearance-based approach is the most popular approach in the research of semantic learning, where semantic knowledge is acquired by interpreting appearance features from sensory data. In [9], O. M. Mozos et al. use geometric features from a planar laser range finder for indoor place classification. This work is extended to incorporate vision features for better and finer classification in [8]. In [11], I. Posner et al. use fused vision and 3D laser data for semantic labelling of urban scenes. Visual features and 2D/3D geometric features are extracted and fed into a hierarchical classifier for scene recognition. Some other appearance-based semantic mapping approaches can be found in [14], [10].

Unlike the appearance-based approaches where semantics is directly learned from sensor readings, the object-based approaches infer the semantic meaning of an environment by checking the occurrence of key objects inside. In [4], C. Galindo et al. infer the semantic type of a room by detecting the typical objects in it. In [16], S. Vasudevan et al. propose to perform place classification using not only object count information, but also the position relationship between objects.

The activity-based approach is to learn the semantic knowledge of an environment based on agent activities in it. Compared to the extensive literature for the appearance-based approach, relatively few are found in this category. In [18], D. F. Wolf et al. build a semantic 2D grid map according to the occupancy of the space by dynamic entities. Activity-related features are used to classify a place into two semantic types, "street" or "sidewalk". In [19], D. Xie et al. present a method to localize functional objects that affect people behavior in surveillance videos.

In this paper, we present a semantic mapping method based on pedestrian activity patterns in the urban road environment. While an environment serves as the space for different agents to conduct different activities, it can be divided into different functional areas, with each area corresponding to certain types of activities. For this reason, we can infer the semantic meaning of an area from its associated activity information. The activity information of a place should be another important dimension of information, together with the metric information and the semantic information. The metric dimension of a place usually describes some geometric shapes or occupancy information, the semantic dimension denotes its meaning, and the activity dimension describes the agent behaviors in it. Our philosophy is that these three dimensions are highly correlated, and can be inferred from each other.

In our specific application, we want to recognize different functional areas for pedestrians in the urban road environment, i.e. "pedestrian path", "entrance/exit", "crossing" and "sidewalk". By observing pedestrian activity over time, the semantic properties of a place can be inferred from the learned motion patterns on them. Without loss of generality, we focus on motion patterns as the key features of pedestrian activity representation. A pedestrian activity model of the environment is first learned, and then 2D grid semantic mapping is performed by classifying the semantic properties of each grid using the learned activity model. Our proposed method is tested through experiments, and has shown good performance. In general, our method can be extended beyond pedestrian activities to vehicles, cyclists, or other agents in the outdoor environment.

The contribution of this paper is clear: to our knowledge, it is the first time to propose the idea of semantic mapping by learning agents' spatial activity models, especially with a mobile platform. The remainder of this paper is organized as follows. Section II gives a brief overview of our system. Section III describes pedestrian activity learning. In section

[1] B. Qin, Z. J. Chong, M. H. Ang Jr. are with the National University of Singapore, Kent Ridge, Singapore {baoxing.qin, chongzj, mpeangh} at nus.edu.sg

[2] T. Bandyopadhyay is with the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia tirtha.bandy at csiro.au

[3] E. Frazzoli and D. Rus are with the Massachusetts Institute of Technology, Cambridge, MA., USA frazzoli at mit.edu, rus at csail.mit.edu

978-1-4799-3684-7/14/$31.00 ©2014 IEEE

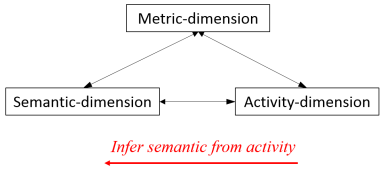


Fig. 1. Correlated multiple dimensions of information

IV, we introduce our algorithm of semantic mapping from pedestrian activity. Experimental results and analysis are presented in Section V. Finally, section VI concludes the paper and discusses future work.

II. System Overview

A. Multi-dimensional Grid Map

In the field of metric mapping, Occupancy Grid Map (OGM) is one of the most popular representation [15]. It represents the environment by evenly spaced grids, with each grid corresponding to a variable of occupancy to be estimated. In this work, we extend the idea of OGM into a multi-dimensional grid map, where each grid has multiple dimensions of information. The multi-dimensional grid map can be formulated as follows: $M = \{m_{ij}|\ 0 \leq i \leq w - 1,\ 0 \leq j \leq h-1\}$, $m_{ij} = (\mathfrak{M}_{ij}, \mathfrak{S}_{ij}, \mathfrak{A}_{ij})^T$. M denotes the map, m_{ij} the grid cell indexed by i and j, which is composed of multiple dimensions of information: metric information $\mathfrak{M}_{ij}$, semantic information $\mathfrak{S}_{ij}$, and activity information $\mathfrak{A}_{ij}$. The width and height of the map are denoted as w and h respectively. These different dimensions of information are correlated, and can be inferred from each other: knowing the metric property of a place will help to infer its semantic meaning, vice versa; the semantic meaning of a place may help robot to infer its normal agent activity, vice versa; etc. In our application, we want to infer the semantic dimension of information from the activity dimension, as shown in Fig. 1.

B. System Framework

In this paper, we want to realize semantic mapping from learning pedestrian activity in the urban road environment, with a mobile platform of an autonomous vehicle. The system framework is illustrated by Fig. 2. Firstly, pedestrians are detected and tracked using on-board sensors, and the collected tracks are then transformed into the global map frame using vehicle localization function. Secondly, track classification and clustering is performed. Thirdly, activity information from moving tracks are registered into the grid map, and then pedestrian activity model is learned. Finally, the semantic information $\mathfrak{S}_{ij}$ is inferred from the learned activity pattern $\mathfrak{A}_{ij}$, together with prior road network information.

III. Pedestrian Activity Learning

This section presents our method of pedestrian activity learning from a mobile platform. We will learn the activity information of each grid $\mathfrak{A}_{ij}$ from collected pedestrian tracks.

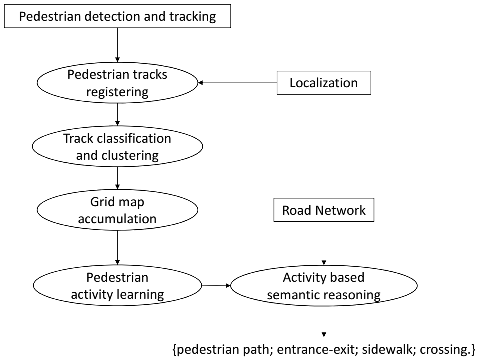

Pedestrian detection and tracking
Pedestrian tracks registering
Localization
Track classification and clustering
Grid map accumulation
Road Network
Pedestrian activity learning
Activity based semantic reasoning
{pedestrian path; entrance-exit; sidewalk; crossing.}

Fig. 2. System Framework

Activity model learning is not a new topic in computer vision community, where researchers have proposed various methods to learn pedestrian motion patterns. Some representative work can be found in [3], [17]. However, most of the algorithms use a stationary camera and assume the observability of complete trajectories, which is not a valid assumption for applications using mobile robots. A. Lookingbill et al. in [6] use a helicopter to identify moving objects on ground and learn their motion patterns. This work shows interesting results and enlightens us about representing motion pattern in the form of grid map. However, it only estimates the motion patterns of grids where moving objects are observed, and it also neglects the relationship between neighboring places. In our work, we use Gaussian Process to learn the activity model of the entire environment, which is able to overcome the above problems.

In this section, we will first discuss the acquirement of pedestrian tracks, then introduce the classification and clustering of tracks, and finally presents the GP-based motion model learning.

A. Pedestrian Detection and Tracking

Pedestrian detection and tracking is one fundamental function for pedestrian activity learning, which is performed using onboard sensors. A laser range finder is used for pedestrian hypothesis generation and tracking, and a camera is used for hypothesis verification. For more details, please refer to our previous work in [13]. The output tracks are sequences of pedestrian positions with time stamps, from which moving speed and direction can also be calculated. While pedestrians are initially detected in the local coordinate of the vehicle, we transform the track information into the global frame of map, using vehicle localization function [1].

B. Track Classification and Clustering

Before using the collected tracks for pedestrian activity learning, track classification and clustering should be applied. Due to the noise in the pedestrian detection and tracking, plus the noise incurred by localization error during track transformation, the motion of some tracks may be very unstable, or they are simply not tracks for pedestrians. In some other cases, static tracks may appear when pedestrians stand still for long time in some places. These tracks are not useful for pedestrian's activity (dynamic) learning, and should be filtered out. A classification process is used to classify the tracks into three types, "moving", "static", and "noisy". The classification is based on several features about the track, such as track length, moving speed, etc. Only "moving" tracks will be used for the activity learning purpose.

Track clustering is performed to cluster heterogeneous tracks into different homogeneous groups. In related researches from computer vision community, pedestrian tracks are usually delicately clustered into multiple groups of high similarity. In our work, however, the mobile platform works in a fairly large area and may collect pedestrians from many heterogeneous motion types. Performing delicate clustering and learning the activity model for each of these types are computationally expensive or even infeasible. On the other hand, since our interest is the activity patterns at individual places, rather than those of the complete pedestrian trajectory spanning in the temporal domain, there is hence no need to perform such clustering and learning.

In fact, from a microscopic view, for an individual place in the urban road environment, there are usually only two dominant motion patterns of pedestrians, which have similar speed but opposite directions. We denote this assumption as the "bidirectional property" of pedestrian activity. While this assumption appears arbitrary at the first glance, it generally holds true for the urban road environment, where pedestrians walk either along or across the road links. This "bidirectional property" simplifies our clustering problem: we cluster the moving tracks scattered over the map into two groups, and only need to guarantee that the activity of each group is consistent at the microscopic grid level.

Our clustering algorithm can be formulated as follows. The set of pedestrian tracks is denoted as $S = \{s_1, \dots, s_m\}$. One track s is a set of position-speed-angle tuples: $s = \{t_1, \dots, t_n\}, t_i =< x_i, y_i, v_i, \theta_i >$, where x_i, y_i are pedestrian positions, v_i the speed, and θ_i the moving direction. The input of the clustering is S, and the output is two clusters of tracks A and B. During the clustering process, each cluster will maintain a set of tuples as its characteristic quality, which is an assembly of the tuples from all its member tracks. The two tuple sets are denoted as α and β respectively.

The similarity between two tuples p, q are defined as:

$$sim_{p,q} = \frac{1 - 2|\theta_P - \theta_q|/\pi}{||x_p - x_q, y_p - y_q|| + const.}$$

The similarity score between a track s and a cluster with tuple set γ is defined as: $SIM_{s,\gamma} = \sum_i^n (max_{p \in \gamma} sim_{t_i,p} + min_{q \in \gamma} sim_{t_i,q})$

During the clustering process, the longest track is first picked out as the seed track for cluster A, and its tuples form the tuple set α. Then the track having the minimum similarity value with α is selected as the seed track for cluster B, whose tuples then form the tuple set β. The track having the highest similarity score with either cluster A or B is assigned to cluster A or B accordingly, until all the tracks are clustered. The pseudo-code of the cluster algorithm can be found in Alg. 1.

```
Input: The set of pedestrian tracks
    $S = \{s_1, \dots, s_m\}$
Output: clusters of tracks A and B
A = B = $\alpha = \beta = \emptyset$;
Find the longest track $s_l$;
Add $s_l$ to A; Add tuples into $\alpha$; Erase $s_l$ from $S$;
Let $s_k = arg\ min_{s_k \in S} SIM_{s_k,\alpha}$;
Add $s_k$ to B; Add the tuples into $\beta$; Erase $s_k$ from $S$;
while $S! = \emptyset$ do
    $score\_A = max_{s_p \in S} SIM_{s_p,\alpha}$;
    $s_p = arg\ max_{s_p \in S} SIM_{s_p,\alpha}$;
    $score\_B = max_{s_q \in S} SIM_{s_q,\beta}$;
    $s_q = arg\ max_{s_q \in S} SIM_{s_q,\beta}$;
    if $score\_A \geq score\_B$ then
        add $s_p$ to A; add tuples into $\alpha$; erase $s_p$ from $S$;
    else
        add $s_q$ to B; add tuples into $\beta$; erase $s_q$ from $S$;
    end
end
return A and B;
```

Algorithm 1: Pseudo-code for track clustering

C. Activity Learning with Gaussian Process

After the classification and clustering process, we get two clusters of moving tracks. The tracks from the same cluster share similar grid-level motion patterns, which are of our interest and need to be learned. We use Gaussian Process (GP) method for this activity learning purpose. To briefly introduce GP, it is a collection of random variables, any finite number of which have (consistent) joint Gaussian distribution. GP can be used to solve both regression and classification problems. Please refer to [13] for more details. We model our activity learning as Gaussian Process Regression (GPR). The set of position-speed-angle tuples for each cluster serves as the observation input, and the prediction output is the information of pedestrian speed v_{ij} and angle θ_{ij}: $\mathfrak{A}_{ij} = \{\bar{v}_{ij}, \sigma^2_{v_{ij}}, \bar{\theta}_{ij}, \sigma^2_{\theta_{ij}}\}^T$.

It should be mentioned that while the speed value can be estimated directly from GPR, it is not suitable to do so for pedestrian moving angle. Unlike a linear variable distributed in $(-\infty, +\infty)$, the angle variable is a circular variable in $[0, 2\pi)$, whose mean and variance are "circular mean" and "circular variance" to be calculated in different ways. For a simple example, the difference between angle 1° and 359° is actually 2°, rather than 358° as calculated in the linear way. According to direction statistics [7], we model the angle distribution as a Projected Normal Distribution, which can be calculated from the bivariate normal distribution of the speed vector $\vec{v} = (v_x, v_y)$. In the activity learning process, three separate GPRs will be trained, with one for the scalar speed

v, and the other two for speed values in x and y directions v_x , v_y. By assuming the independence of v_x and v_y, we can synthesize the bivariate distribution of the speed vector, from which the distribution of moving angle can be calculated.

1) Gaussian Process Regression Model: Let X be the 2-dimension position vector in the map coordinate, $X \in \mathbb{R}^2$, $X = (x, y)$. Let Y be the output value, $Y \in \mathbb{R}$. Our Gaussian Process Regression model is as follows: $Y = F(X)+\xi,\ F \sim GP(m, K),\ \xi \sim N(0, {\sigma_n}^2)$ where $F(X)$ is a function distributed as a GP with mean function m and covariance function K. It can be calculated that the output function Y is also distributed as a GP: $Y \sim GP(m, K+{\sigma_n}^2\sigma_{ii'})$, where $\sigma_{ii'} = 1$ iff $i = i'$. Given a set of training data (X, Y), the posterior distribution for a set of test points X^* is Gaussian distribution: $Y^*|Y \sim N(m(X^*) + K^T(X, X^*)K^{-1}(X, X)(Y - m(X)), K(X^*, X^*) - K^T(X, X^*)K^{-1}(X, X)K(X, X^*))$ In our application, we want to get the posterior distribution for v, v_x, v_y at each test point X_{ij}, where X_{ij} is the position of m_{ij} in the map frame. For this purpose, three separate GPRs are trained, using the tuple set of each cluster as the training data. Zero mean function m and squared exponential covariance function K are used in our GPRs, where $m(X) = 0, K(X, X') = {\sigma_y}^2 \exp \frac{-(X-X')^2}{2l^2}$. The hyperparameters (σ_n, σ_y, l) are learned by maximizing the log-likelihood of the observation in the training data.

2) Projected Normal Distribution of Moving Angles: We use Projected Normal Distribution (PND) to model the probabilistic density function of the pedestrian moving angle. Let $\vec{x}$ be a random two-dimension vector which has a normal distribution $N_2(\mu, \Sigma)$, in which case the angle of $\vec{x}$ is said to have a projected normal (or angular Gaussian) distribution $PN_2(\mu, \Sigma)$. The probabilistic density function of $PN_2(\mu, \Sigma)$ is as follows: $p(\theta; \mu, \Sigma) = \frac{\vartheta(\mu;0,\Sigma)+|\Sigma|^{-\frac{1}{2}}D(\theta)\Phi(D(\theta))\phi(|\Sigma|^{-\frac{1}{2}}(x^T\Sigma^{-1}x)^{-\frac{1}{2}}\mu\wedge x)}{x^T\Sigma^{-1}x}$, where $\vartheta(\mu; 0, \Sigma)$ denotes the value of the probability density function for $N_2(0, \Sigma)$ at point μ, Φ and ϕ denote the probability density function and cumulative density function of $N(0, 1)$, $x = (cos\theta, sin\theta)^T$, $D(\theta) = \frac{\mu^T\Sigma^{-1}x}{(x^T\Sigma^{-1}x)^{-1/2}}$, and $\mu \wedge x = \mu_1 \sin\theta - \mu_2 \cos\theta$ with $\mu = (\mu_1, \mu_2)^T$.

In our application, pedestrian moving direction is actually distributed according to PND. To calculate the distribution, the normal distribution of pedestrian speed vector $\vec{v}$ is used. It is synthetized using the marginalized distribution of v_x and v_y by assuming their independence: $\vec{v} \sim N(diag(\mu_{v_x}, \mu_{v_y}), diag({\sigma_{v_x}}^2, {\sigma_{v_y}}^2))$ With this bivariate normal distribution, the probabilistic density function of moving angle can be calculated. In our semantic reasoning, the circular mean of the distribution $\bar{\theta}_{ij}$ is adopted as pedestrian moving angle, and the circular variance is used to represent the uncertainty of this moving angle $\sigma^2_{\bar{\theta}_{ij}}$. For the detailed definition and calculation of circular mean and variance, please refer to [7].

3) Bidirectional Property of Pedestrian Activity: As discussed in Section 3.2, we classify the collected pedestrian tracks into 2 clusters, and learn their activity models independently. According to real experiments, two learned activity models are actually like a mirror-pair: the moving direction of a place is actually the opposite direction of the other. This leads us to the assumption that pedestrian activity at a place is often "bidirectional", which allows us to learn the activity model of one track cluster, and infer the other via rotating its direction by 180°. In our application, we choose to learn the activity model of the first cluster. Track information from the second cluster is also utilized in the activity learning: the angle values in its activity tuples are added by 180°, and the tuples are used together with the training data from the first cluster.

In the later section of semantic reasoning, we will use the right angle between pedestrian moving direction and road link direction as a feature to infer a place's semantics. Since this angle difference calculated with either of the two activity models is the same, we will use the first model as the deputy for both.

IV. Activity-based Semantic Reasoning

This section introduces our method of activity-based semantic mapping. We want to perform two levels of semantic reasoning, one coarse-level to identify "pedestrian path" (shorthand as "PP"), and one refined-level reasoning to recognize three different types of functional areas from the path, which includes "entrance/exit" (EE), "crossing" (CR), and "sidewalk" (SW). It should be mentioned that these three types of areas are not necessarily mutually exclusive, considering the fact that the same area may serve for different purposes at the same time. To capture the semantic properties at place m_{ij} in the map, a semantic vector of four binary variables is introduced: $\mathfrak{S}_{ij} = (p_{ij}, e_{ij}, c_{ij}, s_{ij})^T$, where
p_{ij}, a binary variable for "path", Λ_p={PP, non-PP}, $p_{ij} \in \Lambda_p$;
e_{ij}, for "entrance/exit", Λ_e={EE, non-EE}, $e_{ij} \in \Lambda_e$;
c_{ij}, for "crossing", Λ_c={CR, non-CR}, $c_{ij} \in \Lambda_c$;
s_{ij}, for "sidewalk", Λ_s={SW, non-SW}, $s_{ij} \in \Lambda_s$;

The input information of the semantic reasoning process is the activity information $\mathfrak{A}_{ij}$, and prior road network information.

A. Pedestrian Path Learning

1) Pedestrian Intensity: In the urban road environment, there are certain explicit or implicit paths that pedestrians can take. The more pedestrians that pass through a certain place, the higher likelihood for it to be part of pedestrian paths. In another word, the pedestrian number counted in one place can be a useful indicator to distinguish its semantic type. Based on this idea, we introduce a measurement "pedestrian intensity" as a feature for pedestrian path classification. The intensity at place m_{ij} is denoted as I_{ij}, which is a function of pedestrian count N_{ij} : $I_{ij} = I_{local_{ij}} \times I_{global_{ij}}$ where $I_{local_{ij}} = N_{i,j}/\underset{a,b}{max}(N_{i+a,j+b}), I_{global_{i,j}} = \frac{1}{1+exp(-N_{ij}+N_{exp})} - \frac{1}{1+exp(N_{exp})}$; Where $a, b \in Z \cap [-l/2, l/2]$. The pedestrian intensity is the multiplication of two factors, the local factor and the global factor, denoted by $I_{local_{ij}}$ and $I_{global_{ij}}$. The local factor is used to normalize the pedestrian count with the maximum values in a $l \times l$ local window. This factor will help to mitigate the problem of data unbalance, which will arise when the observation periods for different areas are too different, leading to the unbalance that pedestrian tracks in some areas are intensively collected while other areas may be overlooked. This local factor has similar effects as the adaptive threshold in image processing, in which it can be used to recover image details when image brightness are unbalanced. The global factor is namely a logistic function of $N_{i,j}$, which increases quickly when $N_{i,j}$ is nearby N_{exp}, while changes slowly when far away. N_{exp} is a constant value chosen as the expected pedestrian count at a "path" place.

2) Classification using Markov Random Field (MRF): Based on pedestrian intensity calculated from the previous step, we use Markov Random Field (MRF) for path classification. MRF is a popular technique in image processing, which can capture the dependency between neighboring pixels and is widely used for image segmentation, restoration and other purposes. For more details please refer to [5].

In this paper, we model our classification problem as a pairwise MRF: Given the intensity data $I = \{I_{i,j}\}$, we want to estimate the "path" semantics of the map $P = \{p_{ij}\}$. Let's assume that $I_{i,j}|p_{ij} \sim N(\mu_{p_{ij}}, \sigma_{p_{ij}})$, where $\mu_{p_{ij}}$ and $\sigma_{p_{ij}}$ can be learned through training data. We get the energy function $U = \sum_{ij}\left(\log(\sqrt{2\pi}\sigma_{p_{ij}})j + \frac{(I_{ij}-\mu_{p_{ij}})^2}{2\sigma^2_{p_{ij}}}\right) + \sum_{i,j,h,k}\beta\delta(p_{ij}, p_{hk})$ where p_{ij} and p_{hk} are "path variables" of neighboring places m_{ij} and m_{hk}, β is a weighting parameter, $\beta \geq 0$. By minimizing this energy function, the optimal classification for pedestrian path can be found, denoted as $\hat{P} = \{\widehat{p_{ij}}\}$.

B. Refined Semantics Learning

After the coarse-level semantic learning for pedestrian path, we want to perform refined semantic reasoning to learn the functional areas in the path, i.e. "entrance/exit", "crossing", and "sidewalk". We use Naive Bayes Classifier (NBC) to learn the semantic variables e_{ij}, c_{ij} and s_{ij}. A Naive Bayes Classifier is a simple probabilistic classifier based on Bayes' theorem assuming independence between features given the class variable. The probability model for a classifier is a conditional model: $p(C|F) = p(C|F_1, \ldots, F_n) = \frac{1}{Z}p(C)\prod_{i=1}^{n} p(F_i|C)$, where C is the class variable, F is the feature set $F = \{F_1, \ldots, F_n\}$, z a normalizer, $p(C)$ the class prior, $p(F_i|C)$ the feature model for F_i given class C, $F_i \in F$.

In our application, three different NBCs are built to classify the three types of functional areas separately, denoted as $p(e_{ij}|F)$, $p(c_{ij}|F)$ and $p(s_{ij}|F)$.We use the similar set of features F for the three NBCs, with different feature models. The set of features used here include "path property" $F_{pp_{ij}}$, "moving direction" $F_{d_{ij}}$, "direction variance" $F_{dv_{ij}}$, and "position" $F_{p_{ij}}$. $F_{ij} = \{F_{pp_{ij}}, F_{d_{ij}}, F_{dv_{ij}}, F_{p_{ij}}\}$.

- $F_{pp_{ij}}$ is a binary feature, which is actually the classification result $\widehat{p_{ij}}$ from the coarse-level "path" classification. The feature model $p(F_{pp_{ij}}|C)$ is designed to carry the idea that if a place is not pedestrian path, it is not likely to be some functional area.
- $F_{d_{ij}}$ is about the angle of pedestrian moving direction. $\bar{\theta}_{ij}$ in $\mathfrak{A}_{ij}$ is chosen as its value. This feature carries the typical motion information at each grid, which is highly related to its semantic meaning.
- $F_{dv_{ij}}$ is about the uncertainty of the learned pedestrian moving angle. $\sigma^2_{\theta_{ij}}$ is chosen as its value. The bigger $F_{dv_{ij}}$ is, the more unreliable about the calculated moving direction $F_{d_{ij}}$.
- $F_{p_{ij}}$ is about a place's relative position to the road network. This feature is introduced with the idea that the functional semantics of a certain place is actually related to its position on the road.

In the rest of this subsection, we will first introduce the prior road information used in the semantic classification process, and then present the classification for each type of semantics. The feature models in different NBCs will be discussed.

1) Prior Road Information: In our previous work [12], we are able to get two kinds of maps for the road network, one binary grid map and one topo-metric map. The binary grid map denotes the binary status of each grid of place, "road" or "non-road". The topo-metric map is a compact representation for the road network, in which road links are represented by fitted splines. We use above two types of road maps to get the required position information in the semantic reasoning process. For example, based on the binary grid map, road boundary information can be retrieved; based on the topo-metric map, the angle of a road link can be calculated from its spline representation; etc.

2) Entrance/Exits $p(e_{ij}|F)$: For urban road environment, pedestrian entrance/exits are where pedestrians enter onto and depart from the road. Due to the bidirectional property of pedestrian motion, people usually use the same pathway for entrance as well as exit into a spatial region. The knowledge of pedestrian entrances and exits in a road network is of vital importance and can help an autonomous vehicle's safe navigation. We use Naive Bayes Classifier to recognize such areas, based on the feature set F_{ij}. The feature models are built as below. (It should be mentioned that above feature model is just a "simplistic abstract" model, which can have different variants in real applications.)

i) $p(F_{pp_{ij}}|e_{ij})$: The entrance/exits (EE) are functional areas of pedestrians, which should only appear on pedestrian path. If an area is "EE", it should be "PP". The infinitesimal ϵ is to avoid degenerate cases. If an area is "non-EE", its possibility to be a "PP" is denoted as k_{ee}, which is approximated by the ratio of extracted "PP" area over the road surface region. It should be mentioned that the same feature models are chosen for the other two semantic properties c_{ij} and s_{ij}, except that different parameters k_{cr} and k_{sw} are used for k_{ee}.

$$
\begin{array}{lll}
p(F_{pp_{ij}} = \text{PP} & |e_{ij} = \text{EE}) & = 1.0 - \epsilon; \\
p(F_{pp_{ij}} = \text{non-PP} & |e_{ij} = \text{EE}) & = \epsilon; \\
p(F_{pp_{ij}} = \text{PP} & |e_{ij} = \text{non-EE}) & = k_{ee}; \\
p(F_{pp_{ij}} = \text{non-PP} & |e_{ij} = \text{non-EE}) & = 1.0 - k_{ee};
\end{array}
$$

ii) $p(F_{d_{ij}}|e_{ij})$: When a pedestrian enters or leaves a road link, its moving direction is usually perpendicular to the road direction. This basic idea is reflected in the feature model, where $F_{d_{ij}}$ is the pedestrian moving direction, rd_{ij} is the direction of the nearest road link calculated from its spline representation, and $\Delta(,)$ is the function to find the right angle between two these two directions.

$$\begin{aligned} p(F_{d_{ij}|e_{ij}} = \text{EE}) \quad &= \frac{4}{\pi}\Delta(F_{d_{ij}}, rd_{ij}); \\ p(F_{d_{ij}|e_{ij}} = \text{non-EE}) \quad &= \frac{2}{\pi}; \end{aligned}$$

iii) $p(F_{dv_{ij}}|e_{ij})$: The feature of angle variance is used to carry the uncertainty of moving direction estimation. This feature model is the same for other two NBCs.

$$p(F_{dv_{ij}|e_{ij}} = \text{EE}) = \frac{2(\max_{i,j} F_{dv_{ij}} - F_{dv_{ij}})}{(\max_{i,j} F_{dv_{ij}} - \min_{i,j} F_{dv_{ij}})^2}$$

$$p(F_{dv_{ij}|e_{ij}} = \text{non-EE}) = \frac{1.0}{\max_{i,j} F_{dv_{ij}} - \min_{i,j} F_{dv_{ij}}}$$

iv) $p(F_{p_{ij}}|e_{ij})$: Pedestrian entrances/exits should appear nearby the road boundary. $F_{p_{ij}}$ denotes a place's distance to the boundary of road, EE_r is a fixed parameter to control the probability. For an area that is "non-EE", the probability density function of $F_{p_{ij}}$ is assumed to be a uniform distribution over $[0, \frac{\text{road\_width}_{ij}}{2.0}]$.

$$\begin{aligned} p(F_{p_{ij}}|e_{ij} = \text{EE}) \quad &= 1.0 - \frac{F_{p_{ij}}}{\text{EE}_r} \\ p(F_{p_{ij}}|e_{ij} = \text{non-EE}) \quad &= \frac{2.0}{\text{road\_width}_{ij}} \end{aligned}$$

With the Naive Bayes Classifier, we can get the "EE" probability of a place. The place with $p(e_{ij} = \text{EE}|\text{F}) > 0.5$ is classified as an EE grid. However, these results are in the format of individual grids, we want to further cluster them into individual EE objects. Gaussian Mixture Model (GMM) is used for the clustering purpose, with which EE grids of places are clustered as EE objects. Each EE object corresponds a 2D position in the map. Bayes Information Criteria (BIC) is used to select the best cluster number. After the clustering, we get a set of EE objects $\xi\xi = \{\text{EE}_1, \ldots, \text{EE}_n\}$.

3) Crossing $p(c_{ij}|F)$: A pedestrian crossing is where pedestrians move across the road. A "crossing" place should be part of the pedestrian "path". From the activity view, pedestrian moving direction should be perpendicular to the road direction. From the position view, the sum of the distances to its two nearest entrance/exits should be around road width. With these ideas, the feature models can be built. With the NBC, we are able to learn a place's semantic property of "crossing". While we can get the classification results directly from NBC, the results may not be smooth in neighbouring areas. In our application, we treat a place's probability of "crossing" as a feature, and input into our MRF framework, to generate better classification results.

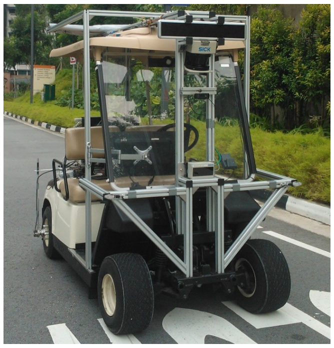


Fig. 3. Vehicle testbed

4) Sidewalk $p(s_{ij}|F)$: A sidewalk is a place where pedestrians walk alongside the road. It should appear near the road boundary, and pedestrian moving direction should be parallel to the road direction. Given these ideas, the feature models of $p(s_{ij}|F)$ can be built. To generate smooth classification results, MRF is used to generate more homogeneous results for s_{ij}, as for c_{ij} discussed previously.

V. Experiments

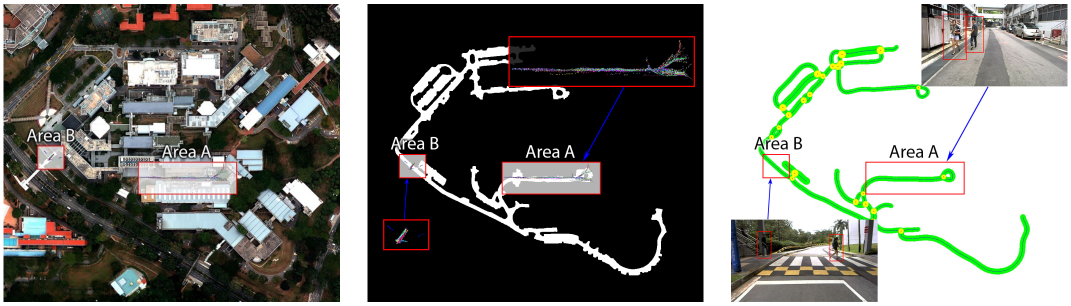
A. Experiment Setup

Our test bed is a Yamaha G22E golf cart with autonomous driving ability [2]. The hardware configuration is shown in Fig. 3. The pedestrian detection and tracking are performed using a 4-layer LIDAR (SICK LD-MRS400001) mounted at the waist height, and a simple webcam above it. Vehicle localization is performed using a tilted-down LIDAR (SICK LMS-291) in the upper front, together with the vehicle's odometry system. Our experiment environment is the Engineering Campus of the National University of Singapore, where pedestrian activities in two typical areas ("Area A" and "Area B") are observed and collected, as shown in Fig 4(a).

A multi-dimensional grid map is built to cover the Engineering Campus, whose size is 1099 × 973 grids, with its resolution 0.5 m/grid. For visualization purposes, we are showing the complete map, but highlight the activity learning and semantic mapping results of the two interesting regions. The size of "Area A" is 338 × 100 grids, and "Area B" is 90 × 90 grids.

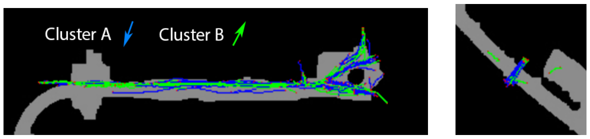
B. Experiment Results

1) Pedestrian Activity Learning: There are 306 tracks collected in our experiment, as shown in Fig. 4. Fig. 4(a) shows the satellite image of our experiment environment, where the two interested areas are highlighted. Collected pedestrian tracks are also overlaid in the picture, drawn in different colors. Fig. 4(b) shows a binary road image from our previous work [26], where white areas are road surface. Pedestrian tracks are overlaid on it. Fig. 4(c) shows the topo-metric graph of the road network, with two sub-images showing the pedestrian detection results in the two areas respectively.



(a) Satellite image (b) Binary road map (c) Topo-metric graph

Fig. 4. Experiment environment and road network information (zoom in when read)



(a) Clustering results at A (b) Clustering results at B

Fig. 5. Track clustering results

The track classification and clustering results are summarized as follows: moving tracks number 205, with 100 in cluster A and 105 in cluster B; static tracks number 10; noisy tracks number 91. It can be seen that static tracks are only a small portion of the track set, meaning that pedestrians in the two surveyed areas are mostly in movement. The relatively large number of "noisy tracks" is due to our strict criteria of track classification, which help us to get reliable moving tracks. For the moving tracks, the similar sizes of the two clusters side-show the "bidirectional property" of pedestrian activity. Fig. 5 visualizes the results of moving track clustering, where cluster A and cluster B are colored in blue and green respectively, and red dots are their end points. Tracks in cluster A generally move from right to left, up to down, where tracks in cluster B takes the opposite direction. The clustering results are checked manually and no errors are found.

Given the results from the track classification and clustering, we try to learn the activity model using Gaussian Process. As discussed in Section 3.3, we only need to learn the activity model in one direction. In this experiment, we learn the activity model in the direction of cluster A. Fig. 6 illustrates the pedestrian moving direction $\bar{\theta}_{ij}$ of the learned activity model. The direction values are shown by red arrows, which is overlaid onto the satellite image for visualization. We can have a glance of the pedestrian motion flow in the environment from this figure.

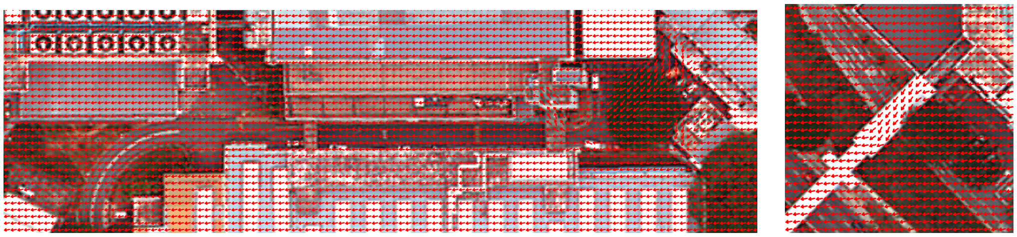
2) Activity-based Semantic Mapping: Together with the road network information, semantic mapping can be performed. Tab. I shows the mapping results for semantic properties of the four types.

For the "path" property, it can be seen that our defined feature of "pedestrian intensity" is able to boost the path trunk which most people take and decline any erratic tracks. The classification results from MRF show complete path and no false positives. For the "entrance/exit" property, the output probability of NBC is shown by a grayscale image, which is overlaid on the satellite image for visualization. The classification results that we get from NBC are individual "EE grids". We use GMM technique to cluster these grids, and recognize the "EE objects". The best cluster number is selected automatically with BIC, and finally we recognize the 7 entrances/exits in the two areas. This result is a perfect result according to our ground truth. For the "crossing" property, we are able to recognize the important crossing area in Area B. However, some grids in Area A (no crossing exists) are misclassified as "crossing". The accuracy for the classification result is 80.3%. This number can be improved to 100% by further filtering out those small pieces of areas according to their size. For the "sidewalk" property, we recognize a long sidewalk in Area A, which has several disconnected pieces at the right end. According to our definition of pedestrian sidewalk, these disconnected pieces do have "sidewalk" property. They can be filtered out according to their sizes if our purpose is to find individual "sidewalk objects" rather than "sidewalk grids". In summary, our activity-based semantic mapping provides promising results. The four types of semantic properties are mapped well in our two survey areas.

VI. Conclusions

In this paper, we propose a novel semantic mapping method based on pedestrian activity in the urban road environment. Pedestrians are detected and tracked using an autonomous vehicle, and the collected track information is used to learn the pedestrian activity model in the environment. Based on the learned pedestrian activity patterns and prior road network information, semantic mapping is performed. Our work is tested through real experiments, and shows promising results.

To detect and track pedestrians using a mobile platform is not as convenient as that using a surveillance camera. In this work, only 306 tracks are collected to test our method. In the future work, we will test our method with more experiments in different road scenarios. Besides pedestrians, there are other equivalently important agents moving on urban roads,



(a) Pedestrian activity at A

(b) Pedestrian activity at B

Fig. 6. Moving direction of activity model (zoom in when read)

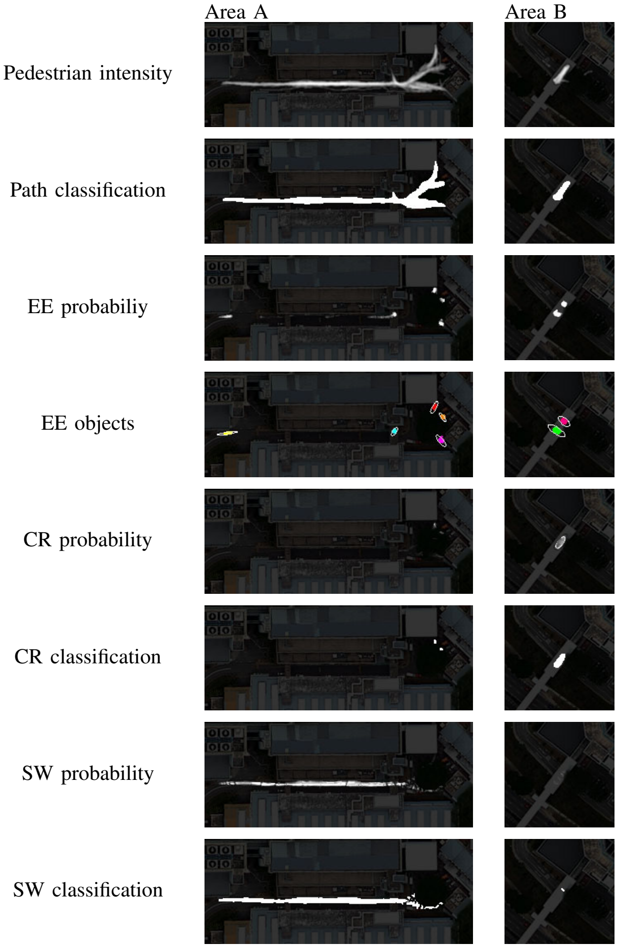


TABLE I
MAPPING RESULTS FOR SEMANTIC PROPERTIES OF THE FOUR TYPES

such vehicles and motorbikes. In our future work, we will extend our method for other types of agents, and learn richer semantic information from their behaviors.

ACKNOWLEDGMENT

This research was supported by the National Research Foundation (NRF) Singapore through the Singapore MIT Alliance for Research and Technology's (FM IRG) research programme, in addition to the partnership with the Defence Science Organisation (DSO). We are grateful for their support.

REFERENCES

[1] Z. J. Chong, B. Qin, T. Bandyopadhyay, M. Ang, E. Frazzoli, and D. Rus, "Synthetic 2d LIDAR for precise vehicle localization in 3d urban environment," in *IEEE International Conference on Robotics and Automation*, Karlsruhe, Germany, May 2013.

[2] Z. J. Chong, B. Qin, T. Bandyopadhyay, T. Wongpiromsarn, E. S. Rankin, M. H. Ang Jr., E. Frazzoli, D. Rus, D. Hsu, and K. H. Low, "Autonomous Personal Vehicle for the First- and Last-Mile Transportation Services," in *IEEE International Conference on Robotics, Automation and Mechatronics (RAM)*, 2011.

[3] D. Ellis, E. Sommerlade, and I. Reid, "Modelling pedestrian trajectory patterns with gaussian processes," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1229–1234.

[4] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J.-A. Fernandez-Madrigal, and J. González, "Multi-hierarchical semantic maps for mobile robotics," in *Intelligent Robots and Systems, IEEE/RSJ International Conference on (IROS)*. IEEE, 2005.

[5] R. Kindermann, J. L. Snell, *et al.*, *Markov random fields and their applications*.

[6] A. Lookingbill, D. Lieb, D. Stavens, and S. Thrun, "Learning activity-based ground models from a moving helicopter platform," in *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*. IEEE, 2005, pp. 3948–3953.

[7] K. V. Mardia and P. E. Jupp, *Directional statistics*. Wiley. com, 2009, vol. 494.

[8] O. Martinez Mozos, A. Rottmann, R. Triebel, P. Jensfelt, W. Burgard, *et al.*, "Semantic labeling of places using information extracted from laser and vision sensor data," 2006.

[9] O. Mozos, C. Stachniss, and W. Burgard, "Supervised learning of places from range data using adaboost," in *Robotics and Automation, 2005. ICRA. Proceedings of the IEEE International Conference on*.

[10] A. Nüchter and J. Hertzberg, "Towards semantic maps for mobile robots," *Robotics and Autonomous Systems*, 2008.

[11] I. Posner, D. Schroeter, and P. Newman, "Online generation of scene descriptions in urban environments," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 901–914, 2008.

[12] B. Qin, Z. J. Chong, T. Bandyopadhyay, and M. H. Ang Jr., "Metric Mapping and Topo-metric Graph Learning of Urban Road Network," in *IEEE International Conference on Robotics, Automation and Mechatronics (RAM)*, 2013.

[13] C. E. Rasmussen, "Gaussian processes for machine learning," 2006.

[14] S. Sengupta, P. Sturgess, P. H. Torr, *et al.*, "Automatic dense visual semantic mapping from street-level imagery," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*.

[15] S. Thrun, W. Burgard, and D. Fox, *Probabilistic robotics*. MIT Press, 2005.

[16] S. Vasudevan and R. Siegwart, "Bayesian space conceptualization and place classification for semantic maps in mobile robotics," *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 522–537, 2008.

[17] X. Wang, K. T. Ma, G.-W. Ng, and W. E. L. Grimson, "Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models," *International journal of computer vision*, 2011.

[18] D. F. Wolf and G. S. Sukhatme, "Semantic mapping using mobile robots," *Robotics, IEEE Transactions on*, vol. 24, no. 2, pp. 245–258, 2008.

[19] D. Xie, S. Todorovic, and S.-C. Zhu, "Inferring "dark matter" and "dark energy" from videos," in *The IEEE International Conference on Computer Vision (ICCV)*, 2013.