Proceedings of the IEEE ICRA 2009
Workshop on People Detection and Tracking
Kobe, Japan, May 2009

Spatially Grounded Multi-Hypothesis Tracking of People

Matthias Luber Gian Diego Tipaldi Kai O. Arras

***Abstract*— People tracking is an important yet challenging task for mobile robots operating in populated environments and interacting with humans. What makes this problem difficult is that human behavior is complex and hard to predict. However, motion of people, the rate at which people appear and where they appear are not random but strongly place-dependent and follow patterns that are engendered by the environment. In this paper we make use of such information for the purpose of people tracking. Concretely, we learn a probabilistic representation, called *spatial affordance map*, to spatially ground activity events acquired by observing people in the environment. This representation is a non-homogeneous spatial Poisson process for which we derive expressions for life-long Bayesian learning. We show how the spatial affordance map can be used to compute refined probability distributions over hypotheses in a multi-hypothesis tracker and to make better, place-dependent predictions of human motion. In experiments with real data from a laser range finder, we demonstrate how both extensions lead to more accurate tracking behavior. The system runs in real-time on a typical desktop computer.**

I. Introduction

As robots enter more domains in which they interact and cooperate closely with humans, people tracking is becoming a key technology for several areas in robotics such as human-robot interaction, intelligent cars or human activity understanding.

In this paper we pursue the approach to learn and represent human spatial behavior for improved people tracking. Human activity is strongly place-dependent. By learning a spatial model that represents activity events in a global reference frame and on large time scales, the robot acquires place-dependent priors on human behavior. As we will demonstrate, such priors can be used to better hypothesize about the state of the world (that is, the state of people in the world), and to make place-dependent predictions of human motion that better reflect how people are using space. Concretely, we propose a non-homogeneous spatial Poisson process to represent the spatially varying distribution over relevant human activity events for people tracking. The representation, called *spatial affordance map*, holds space-dependent Poisson rates for the occurrence of track events such as creation, confirmation or false alarm. The map is then incorporated into a multi-hypothesis tracking framework using data from a laser range finder.

All authors are with the Social Robotics Lab, Department of Computer Science, University of Freiburg, Germany {luber,tipaldi,arras}@informatik.uni-freiburg.de.

In most related work on laser-based people tracking [1], [2], [3], [4], [5], [6], [7], a person is represented as a single state that encodes torso position and velocities. People are extracted from range data as single blobs or found by merging nearby point clusters that correspond to legs. The problem of people tracking has also been addressed as a leg tracking problem [8], [9], [10] where people are represented by the states of two legs, either in a single augmented state [9] or as a high-level track to which two low-level leg tracks are associated [8], [10].

Different tracking and data association approaches have been used for laser-based people tracking. The nearest neighbor filter and variations thereof are typically employed in earlier works [1], [2], [3]. A sample-based joint probabilistic data association filter (JPDAF) has been presented in Schulz *et al.* [4] and adopted by Topp *et al.* [5]. The Multi-hypothesis tracking (MHT) approach according to Reid [11] and Cox *et al.* [12] has been used in [8], [7], [10]. What makes the MHT an attractive choice is that it belongs to the most general data association techniques. The method generates joint compatible assignments, integrates them over time, and is able to deal with track creation, confirmation, occlusion, and deletion events in a probabilistically consistent way. Other multi-target data association techniques such as the global nearest neighbor filter, the track splitting filter or the JPDAF are suboptimal in nature as they simplify the problem in one or the other way [13], [14]. For this reasons, the MHT has become a widely accepted tool in the target tracking community [14].

The MHT framework assumes that new track and false alarm events are uniformly distributed in the sensor field of view with fixed Poisson rates. This assumption is justified in settings for which the approach has been originally developed (using, e.g., radar or underwater sonar). However, in the context of people tracking with vision or laser these models are overly simplified. Particularly since people do not use environments randomly but move, appear and disappear at specific locations that correspond, for instance, to doors, entrances, or convex corners. Further, false alarms are more likely to arise in areas with cluttered backgrounds rather than in open spaces. In this paper, we extend the MHT approach by incorporating learned distributions over track interpretation events that serve as domain knowledge to the system to better hypothesize about the state of the world.

For motion prediction of people, most researchers employ the Brownian motion model and the constant velocity motion model. The former makes no assumptions

about the target dynamics, the latter assumes linear target motion. Better motion models for people tracking have been proposed by Bruce and Gordon [15] and Liao *et al.* [16].

In [15], the robot learns goal locations in an environment from people trajectories obtained by a laser-based tracker. Goals are found as end points of clustered trajectories. Human motion is then predicted along paths that a planner generates from the location of people being tracked to the goal locations. The performance of the tracker was improved in comparison to a Brownian motion model. Liao *et al.* [16] extract a Voronoi graph from a map of the environment and represent the state of people being on edges of that graph. This allows them to predict motion of people along the edges that follow the topological shape of the environment.

With maneuvering targets, a single model can be insufficient to represent the target's motion. Multiple model based approaches in which different models run in parallel and describe different aspects of the target behavior are a widely accepted technique to deal with maneuvering targets, in particular the Interacting Multiple Model (IMM) algorithm [17]. Different target motion models are also studied by Kwok and Fox [18]. The approach is based on a Rao-Blackwellized particle filter to model the potential interactions between a target and its environment. The authors define a discrete set of different target motion models from which the filter draws samples. Then, conditioned on the model, the target is tracked using Kalman filters.

Our approach extends prior work in two aspects, learning and place-dependent motion prediction. Opposed to [16], [18] and IMM related methods, we do not rely on predefined motion models but apply learning for this task in order to acquire place-dependent models. In [16], the positions of people is projected onto a Voronoi graph which is a topologically correct but metrically poor model for human motion. While sufficient for the purpose of their work, there is no insight why people should move on a Voronoi set, particularly in open spaces whose topology is less well defined. Our approach, by contrast, tracks the actual position of people and predicts their motion according to metric, place-dependent models. Opposed to [15] where motion prediction is done along paths that a planner plans to a set of goal locations, our learning approach predicts motion along the trajectories that people are actually following.

The paper is structured as follows: the next section gives an overview of the people tracker that will later be extended. Section III introduces the theory of the spatial affordance map and expressions for learning its parameters. Section IV describes how the spatial affordance map can be used to compute refined probability distributions over hypotheses, while section V contains the theory for the place-dependent motion model. Section VI presents the experimental results followed by the conclusions in section VII.

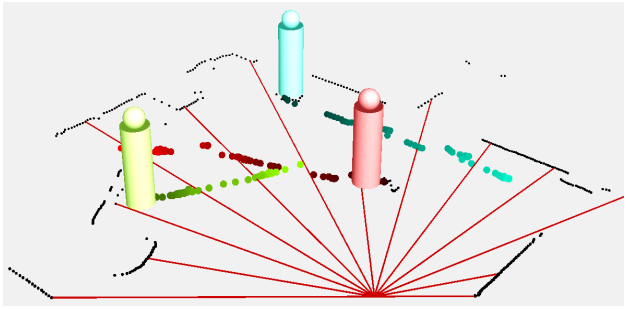

Fig. 1. An example scene from experiment 2 (frame 185) where three people are being tracked.

II. Multi-Hypothesis Tracking of People

For people tracking, we pursue a Multi-Hypothesis Tracking (MHT) approach described in Arras *et al.* [10] based on the original MHT by Reid [11] and Cox and Hingorani [12]. As we will use the tracker to learn the spatial affordance map described hereafter, we give a short outline. Sections IV and V, where the approach will be extended, contains the technical details.

Summarizing, the MHT algorithm hypothesizes about the state of the world by considering all statistically feasible assignments between measurements and tracks and all possible interpretations of measurements as false alarms or new track and tracks as matched, occluded or obsolete. A hypothesis Ω_i^t is one possible set of assignments and interpretations at time t.

For learning the spatial affordance map, the hypothesis with maximal probability Ω_{best}^t at time t is chosen to produce the track events that subsequently serve as observations for the map. In case of a sensor mounted on a mobile platform, we assume the existence of a metric map of the environment and the ability of the robot to self-localize. Observations are then transformed from local, robot-centric coordinates into the world reference frame of the map.

III. Spatial Affordance Map

The spatial affordance map is a non-homogeneous spatial Poisson process. This section describes the theory and how learning is implemented in this application of a Poisson process.

A Poisson distribution is a discrete distribution to compute the probability of a certain number of events given an expected average number of events over time or space. The parameter of the distribution is the positive real number λ, the rate at which events occur per time or volume units. As we are interesting in modeling events that occur randomly in time, the Poisson distribution is a natural choice.

Based on the assumption that events in time occur independently of one another, a *Poisson process* can deal with distributions of time intervals between events. Concretely, let $N(t)$ be a discrete random variable to represent the number of events occurring up to time t

with rate λ. Then we have that $N(t)$ follows a Poisson distribution with parameter λt

$$P(N(t) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!} \qquad k = 0, 1, \ldots \tag{1}$$

In general, the rate parameter may change over time. In this case, the generalized rate function is given as $\lambda(t)$ and the expected number of events between time a and b is

$$\lambda_{a,b} = \int_a^b \lambda(t)\, dt. \tag{2}$$

A homogeneous Poisson process is a special case of a non-homogeneous process with constant rate $\lambda(t) = \lambda$.

The *spatial* Poisson process introduces a spatial dependency on the rate function given as $\lambda(\vec{x}, t)$ with $\vec{x} \in X$ where X is a vector space such as $\mathbb{R}^2$ or $\mathbb{R}^3$. For any subset $S \subset X$ of finite extent (e.g. a spatial region), the number of events occurring inside this region can be modeled as a Poisson process with associated rate function $\lambda_S(t)$ such that

$$\lambda_S(t) = \int_S \lambda(\vec{x}, t)\, d\vec{x}. \tag{3}$$

In the case that this generalized rate function is a separable function of time and space, we have:

$$\lambda(\vec{x}, t) = f(\vec{x})\lambda(t) \tag{4}$$

for some function $f(\vec{x})$ for which we can demand

$$\int_X f(\vec{x})\, d\vec{x} = 1 \tag{5}$$

without loss of generality. This particular decomposition allows us to decouple the occurrence of events between time and space. Given Eq. 5, $\lambda(t)$ defines the occurrence rate of events, while $f(\vec{x})$ can be interpreted as a probability distribution on where the event occurs in space.

Learning the spatio-temporal distribution of events in an environment is equivalent to learn the generalized rate function $\lambda(\vec{x}, t)$. However, learning the full continuous function is a highly expensive process. For this reason, we approximate the non-homogeneous spatial Poisson process with a piecewise homogeneous one. The approximation is performed by discretizing the environment into a bidimensional grid, where each cell represents a local homogeneous Poisson process with a fixed rate over time,

$$P_{ij}(k) = \frac{e^{-\lambda_{ij}}(\lambda_{ij})^k}{k!} \qquad k = 0, 1, \ldots \tag{6}$$

where λ_{ij} is assumed to be constant over time. Finally, the spatial affordance map is the generalized rate function $\lambda(\vec{x}, t)$ using a grid approximation,

$$\lambda(\vec{x}, t) \simeq \sum_{(i,j)\in X} \lambda_{ij} \mathbf{1}_{ij}(\vec{x}) \tag{7}$$

with $\mathbf{1}_{ij}(\vec{x})$ being the indicator function defined as

$$\mathbf{1}_{ij}(x) = \begin{cases} 1 & \text{if } x \in \text{cell}_{ij}, \\ 0 & \text{if } x \notin \text{cell}_{ij}. \end{cases} \tag{8}$$

The type of approximation is not imperative and goes without loss of generality. Other space tessellation techniques such as graphs, quadtrees or arbitrary regions of homogeneous Poisson rates can equally be used. Subdivision of space into regions of fixed Poisson rates has the property that the preferable decomposition in Eq. 4 holds.

Each type of human activity event can be used to learn its own probability distribution in the map. We can therefore think of the map as a representation with multiple layers, one for every type of event. For the purpose of this paper, the map has three layers, one for new tracks, for matched tracks and for false alarms. The first layer represents the distribution and rates of people appearing in the environment. The second layer can be considered a space usage probability and contains a walkable area map of the environment. The false alarm layer represents the place-dependent reliability of the detector.

A. Learning

In this section we show how to learn the parameter of a single cell in our grid from a sequence $K_{1..n}$ of n observations $k_i \in \{0, 1\}$. We use Bayesian inference for parameter learning, since the Bayesian approach can provide information on cells via a prior distribution. We model the parameter λ using a Gamma distribution, as it is the conjugate prior of the Poisson distribution. Let λ be distributed according to the Gamma density, $\lambda \sim \text{Gamma}(\alpha, \beta)$, parametrized by the two parameters α and β,

$$\text{Gamma}(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\,\lambda} \qquad \text{for } \lambda > 0. \tag{9}$$

Then, learning the rate parameter λ consists in estimating the parameters of a Gamma distribution. At discrete time index i, the posterior probability of λ_i according to Bayes' rule is computed as

$$P(\lambda_i | K_{1..i}) \sim P(k_i | \lambda_{i-1})\, P(\lambda_{i-1}) \tag{10}$$

with $P(\lambda_{i-1}) = \text{Gamma}(\alpha_{i-1}, \beta_{i-1})$ being the prior and $P(k_i|\lambda_{i-1}) = P(k_i)$ from Eq. 6 the likelihood. Then by substitution, it can be shown that the update rules for the parameters are

$$\alpha_i = \alpha_{i-1} + k_i \qquad \beta_i = \beta_{i-1} + 1. \tag{11}$$

The posterior mean of the rate parameter in a single cell is finally obtained as the expected value of the Gamma,

$$\widehat{\lambda}_{\text{Bayesian}} = \mathbb{E}[\lambda] = \frac{\alpha}{\beta} = \frac{\#\text{positive events} + 1}{\#\text{observations} + 1}\,. \tag{12}$$

For $i = 0$ the quasi uniform Gamma prior for $\alpha = 1$, $\beta = 1$ is taken. The advantages of the Bayesian estimator are that it provides a variance estimate which is a measure of confidence of the mean and that it allows to properly initialize never observed cells.

Given the learned rates we can estimate the space distribution of the various events. This distribution is

obtained from the rate function of our spatial affordance map $\lambda(\vec{x}, t)$. While this estimation is hard in the general setting of a non-homogeneous spatial Poisson process, it becomes easy to compute if the separability property of Eq. 4 holds[1]. In this case, the pdf, $f(\vec{x})$, is obtained by

$$f(\vec{x}) = \frac{\lambda(\vec{x}, t)}{\lambda(t)} \tag{13}$$

where $\lambda(\vec{x}, t)$ is the spatial affordance map. The nominator, $\lambda(t)$, can be obtained from the map by substituting the expression for $f(\vec{x})$ into the constraint defined in Eq. 5. Hence,

$$\lambda(t) = \int_X \lambda(\vec{x}, t)\, d\vec{x}. \tag{14}$$

In our grid, those quantities are computed as

$$f(\vec{x}) \quad = \quad \frac{\sum_{(i,j)\in X} \lambda_{ij} \mathbf{1}_{ij}(\vec{x})}{\sum_{(i,j)\in X} \lambda_{ij}}. \tag{15}$$

In case of several layers in the map, each layer contains the distribution $f(\vec{x})$ of the respective type of events. Note that learning in the spatial affordance map is simply realized by counting in a grid. This makes life-long learning particularly straightforward as new information can be added at any time by one or multiple robots.

Figure 2 shows two layers of the spatial affordance map of our laboratory, learned during a first experiment. The picture on the left shows the space usage distribution of the environment. The modes in this distribution correspond to often used places and have the meaning of goal locations in that room (two desks and a sofa). On the right, the distribution of new tracks is depicted whose peaks denote locations where people appear (doors). The reason for the peaks at other locations than the doors is that when subjects use an object (sit on a chair, lie on the sofa), they cause a track loss. When they reenter space, they are detected again as new tracks.

IV. MHT With Spatial Information

The Multi-Hypothesis Tracking approach has its roots in the target tracking community and was designed for sensors such as radar or underwater sonar. When employed with data from a mobile platform with cameras or laser range finders, it is questionable if the same statistical assumptions hold. The MHT assumes a Poisson distribution for the occurrences of new tracks and false alarms over time and a uniform probability of these events over space within the sensor field of view V. While this is a valid assumption for a radar aimed upwards into the sky, this is unrealistic for people being tracked by a mobile robot. The arrival of people is well modeled by a Poisson distribution but is clearly non-uniform over space. People typically appear and disappear at specific locations that correspond, for instance, to doors, entrances, or convex corners.

[1]Note that for a non-separable rate function, the Poisson process can model places whose importance changes over time.

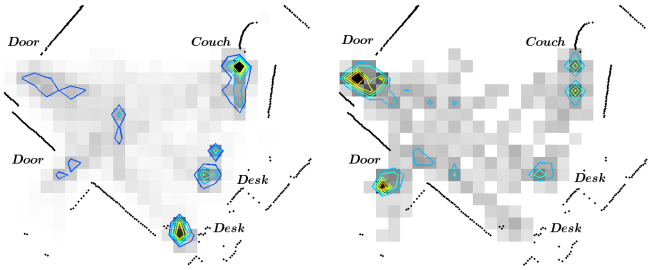

Fig. 2. Spatial affordance map of the laboratory in experiment 1. The probability distribution of matched track events is shown on the left, the distribution of new track events is shown on the right. The marked locations in each distribution (extracted with a peak finder and visualized by contours of equal probability) have different meanings. While on the left they correspond to places that are often used by people (two desks and a sofa), the maxima of the new track distribution (right) denote locations where people appear (two doors and a sofa).

It is exactly this information that the spatial affordance map holds. We can therefore seamlessly extend the MHT approach with the learned Poisson rates for the arrival events of people and learned location statistics for new tracks and false alarms.

At time t, each possible set of assignments and interpretations forms a hypothesis Ω_i^t. Let $Z(t) = \{z_i(t)\}_{i=1}^{m_t}$ be the set of m_t measurements which in our case is the set of detected people in the laser data. For detection, we use a learned classifier based on a collection of boosted features [19]. Let further $\psi_i(t)$ denote a set of assignments which associates predicted tracks to measurements in $Z(t)$ and let Z^t be the set of all measurements up to time t. Starting from a hypothesis of the previous time step, called a parent hypothesis $\Omega_{p(i)}^{t-1}$, and a new set $Z(t)$, there are many possible assignment sets $\psi(t)$, each giving birth to a child hypothesis that branches off the parent. This makes up an exponentially growing hypothesis tree. For a real-time implementation, the growing tree needs to be pruned. To guide the pruning, each hypothesis receives a probability, recursively calculated as the product of a normalizer η, a measurement likelihood, an assignment set probability and the parent hypothesis probability [11],

$$\begin{aligned} p(\Omega_l^t \mid Z^t) \quad &= \quad \eta \cdot p(Z(t) \mid \psi_i(t), \Omega_{p(i)}^{t-1}) \\ &\quad\; p(\psi_i(t) \mid \Omega_{p(i)}^{t-1}, Z^{t-1}) \cdot p(\Omega_{p(i)}^{t-1} \mid Z^{t-1}). \end{aligned} \tag{16}$$

While the last term is known from the previous iteration, the two expressions that will be affected by our extension are the measurement likelihood and the assignment set probability.

For the measurement likelihood, we assume that a measurement $z_i(t)$ associated to a track $\mathbf{x}_j$ has a Gaussian pdf centered on the measurement prediction $\hat{z}_j(t)$ with innovation covariance matrix $S_{ij}(t)$, $\mathcal{N}(z_i(t)) := \mathcal{N}(z_i(t); \hat{z}_j(t), S_{ij}(t))$. The regular MHT now assumes that the pdf of a measurement belonging to a new track or false alarm is uniform in V, the sensor field of view,

with probability V^{-1}. Thus

$$p(Z(t) \mid \psi_i(t), \Omega_{p(i)}^{t-1}) = V^{-(N_F+N_N)} \cdot \prod_{i=1}^{m_t} \mathcal{N}(z_i(t))^{\delta_i} \quad (17)$$

with N_F and N_N being the number of measurements labeled as false alarms and new tracks respectively. δ_i is an indicator variable being 1 if measurement i has been associated to a track, and 0 otherwise.

Given the spatial affordance map, the term changes as follows. The probability of new tracks V^{-1} can now be replaced by

$$p_N(\vec{x}) = \frac{\lambda_N(\vec{x}, t)}{\lambda_N(t)} = \frac{\lambda_N(\vec{x}, t)}{\int_V \lambda_N(\vec{x}, t)\, d\vec{x}} \quad (18)$$

where $\lambda_N(\vec{x}, t)$ is the learned Poisson rate of new tracks in the map and $\vec{x}$ the position of measurement $z_i'(t)$ transformed into global coordinates. The same derivation applies for false alarms. Given our grid, Eq. 18 becomes

$$p_N(\vec{x}) = \frac{\lambda_N(z_i'(t), t)}{\sum_{(i,j)\in V} \lambda_{ij,N}}. \quad (19)$$

The probability of false alarms $p_F(\vec{x})$ is calculated in the same way using the learned Poisson rate of false alarms $\lambda_F(\vec{x}, t)$ in the map.

The original expression for the assignment set probability can be shown to be [10]

$$\begin{aligned} p(\psi_i(t) \mid \Omega_{p(i)}^{t-1}, Z^{t-1}) &= \eta' \cdot p_M^{N_M} \cdot p_O^{N_O} \cdot p_D^{N_D} \quad (20) \\ &\quad \lambda_N^{N_N} \cdot \lambda_F^{N_F} \cdot V^{(N_F+N_N)} \end{aligned}$$

where N_M, N_O, and N_D are the number of matched, occluded and deleted tracks, respectively. The parameters p_M, p_O, and p_D denote the probability of matching, occlusion and deletion that are subject to $p_M + p_O + p_D = 1$. The regular MHT now assumes that the number of new tracks N_N and false alarms N_F both follow a fixed rate Poisson distribution with expected number of occurrences $\lambda_N V$ and $\lambda_F V$ in the observation volume V.

Given the spatial affordance map, they can be replaced by rates from the learned spatial Poisson process with rate functions $\lambda_N(t)$ and $\lambda_F(t)$ respectively.

Substituting the modified terms back into Eq. 16 makes, like in the original approach, that many terms cancel out leading to an easy-to-implement expression for a hypothesis probability

$$\begin{aligned} p(\Omega_l^t \mid Z^t) = \eta'' \cdot p_M^{N_M} \cdot p_O^{N_O} \cdot p_D^{N_D} \cdot \prod_{i=1}^{m_t} [\mathcal{N}(z_i(t))^{\delta_i} \quad (21) \\ \lambda_N(z_i'(t), t)^{\kappa_i} \cdot \lambda_F(z_i'(t), t)^{\phi_i}] \cdot p(\Omega_{p(i)}^{t-1} \mid Z^{t-1}) \end{aligned}$$

with δ_i and κ_i being indicator variables whether a track is matched to a measurement or new, respectively, and ϕ_i indicating if a measurement is declared to be a false alarm.

The insight of this extension of the MHT is that we replace fixed parameters by learned distributions. This kind of domain knowledge helps the tracker to better interpret measurements and tracks, leading to refined probability distributions over hypotheses at the same run-time costs.

V. Place-Dependent Motion Models

Tracking algorithms rely on the predict-update cycle, where a motion model predicts the future target position which is then validated by an observation in the update phase. Without validation, caused, for instance, by the target being hidden during an occlusion event, the state evolves blindly following only the prediction model. Good motion models are especially important for people tracking as people typically undergo lengthy occlusion events during interaction with each other or with the environment.

As motion of people is hard to predict, having a precise model is difficult. People can abruptly stop, turn back, left or right, make a step sideways or accelerate suddenly. However, motion of people is not random. In particular, it follows patterns that are strongly place-dependent. They, for instance, turn around convex corners, avoid static obstacles, stop in front of doors and do not go through walls. Clearly, the Brownian and the constant velocity motion model are unable to capture the complexity of these movements and even higher-order models would be a very approximate choice.

For this reason, we extend the constant velocity motion assumption with a place-dependent model derived from the learned space usage distribution in the spatial affordance map. Let $\mathbf{x}_t = (\ x_t \quad y_t \quad \dot{x}_t \quad \dot{y}_t\ )^T$ be the state of a track at time t and Σ_t its covariance estimate. The motion model $p(x_t|x_{t-1})$ is then defined as

$$p(x_t|x_{t-1}) = \mathcal{N}(x_t; F\, x_{t-1}, F\, \Sigma_{t-1}\, F^T + Q) \quad (22)$$

with F being the state transition matrix. The entries in Q represent the acceleration capability of a human. We extend this model by considering how the distribution of the state at a generic time t is influenced by the previous state and the map. This distribution is approximated by the following factorization

$$p(x_t|x_{t-1}, m) \simeq p(x_t|x_{t-1}) \cdot p(x_t|m) \quad (23)$$

where m is the spatial affordance map and $p(x_t|m) = f(x)$ denotes the space usage probability of the portion of the environment occupied by x_t, as defined by Eq. 15.

A closed form estimation of this distribution does not exist since the map contains a general density, poorly described by a parametric distribution. We therefore follow a sampling approach and use a particle filter to address this estimation problem. The particle filter is a sequential Monte Carlo technique based on the importance sampling principle. In practice, it represents a target distribution in form of a set of weighted samples

$$p(x_t|x_{t-1}, m) \simeq \sum_i w^{(i)} \delta_{x_t^{(i)}}(x_t). \quad (24)$$

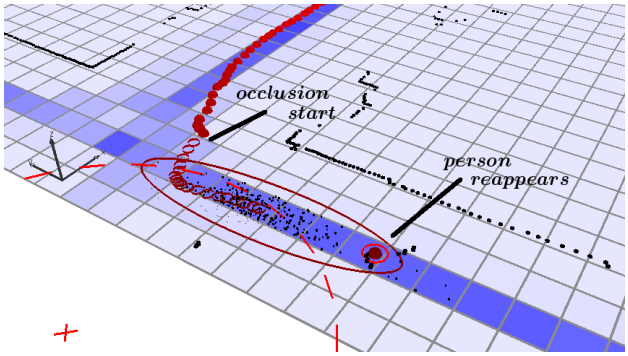

Fig. 3. Trajectory of a person in experiment 2 taking a left turn during an occlusion event. Predictions from a constant velocity motion model (dashed ellipse) and the new model (solid ellipse) are shown. The background grid (in blue) shows the learned space usage distribution of the spatial affordance map. The small black dots are the weighted samples of the place-dependent motion model. The model is able to predict the target "around the corner" yielding much better motion predictions in this type of situations.

where $\delta_{x_t^{(i)}}(x_t)$ is the impulse function centered in $x_t^{(i)}$. Sampling directly from that distribution is not possible so the algorithm first computes samples from a so called proposal distribution, π. The algorithm, then, computes the importance weight related to the i-th sample that takes into account the mismatch among the target distribution τ and the proposal distribution $w = \frac{\tau}{\pi}$. The weights are then normalized such that $\sum w = 1$.

In our case, we take the constant velocity model to derive the proposal π. The importance weights are then represented by the space usage probability

$$w^{(i)} = \frac{p(x_t|x_{t-1}, m)}{p(x_t^{(i)}|_{t-1})} = p(x_t^{(i)}|m). \tag{25}$$

The new motion model has now the form of a weighted sample set. Since we are using Kalman filters for tracking, the first two moments of this distribution is estimated by

$$\hat{\mu} = \sum_i w^{(i)} x_t^{(i)} \tag{26}$$

$$\hat{\Sigma} = \sum_i w^{(i)} (\hat{\mu} - x_t^{(i)})(\hat{\mu} - x_t^{(i)})^T. \tag{27}$$

The target is then predicted using $\hat{\mu}$ as the state prediction with associated covariance $\hat{\Sigma}$. Obviously, the last step is not needed when using particle filters for tracking.

An example situation that exemplifies how this motion model works is shown in Figure 3. A person that takes a left turn in a hallway is tracked over a lengthy occlusion event. The constant velocity motion model (dashed ellipse) predicts the target into a wall and outside the walkable area of the environment. The place-dependent model (solid ellipse) is able to follow the left turn with a state covariance in the shape of the hallway. In other words, the model predicts the target "around the corner". The tracker with the constant velocity motion loses track as the reappearing person is outside the validation gate (shown as 95% ellipses).

VI. Experiments

For the experiments we collected two data sets, one in a laboratory (experiment 1, Figure 4) and one in an office building (experiment 2, Figure 6). As sensors we used a fixed Sick laser scanner with an angular resolution of 0.5 degree.

The spatial affordance maps were trained based on the tracker described in [10], the grid cells were chosen to be 30 cm in size. The parameters of the tracker have been learned from a training data set with 28 tracks over 889 frames. All data associations including occlusions have been hand-labeled. This led to a matching probability $p_M = 0.515$, an occlusion probability $p_O = 0.472$, a deletion probability $p_D = 0.013$, a fixed Poisson rate for new tracks $\lambda_N = 0.033$ and a fixed Poisson rate for false alarms as $\lambda_F = 0.0011$. The rates have been estimated using the Bayesian approach in Eq. 12.

The implementation of our system runs in real-time on a 2.8 GHz quad-core CPU. The cycle time of a typical setting with $N_{Hyp} = 50$, 500 samples for the particle filter, and up to eight parallel tracks is around 12 Hz when sensor data are immediately available.

A. MHT with Spatial Information

The original MHT is compared to the approach using the spatial affordance map on the data set from the laboratory over 4588 frames and with a total number of 130 people entering and leaving the sensor field of view. The ground truth has been determined by manual inspection. For the comparison we count the total number of tracks that are created by the current best hypotheses of the two tracking methods. This value is indication of the tracking accuracy, especially of the ability to deal with track occlusion. We use a pruning strategy which limits the maximum number of hypotheses at every step to N_{Hyp} (the multi-parent variant of the pruning algorithm proposed by Murty [20]). In order to show the evolution of the error as a function of N_{Hyp}, the computational effort, N_{Hyp} is varied from 1 to 50. The results are shown in Figure 5.

The result shows a significant improvement of the extended MHT over the regular approach. The explanation is given by an example. As can be seen in Figure 2 right, few new track events have been observed in the center of the room. If at such a place a track occlusion occurs (e.g. from another person), hypotheses that interpret this as an obsolete track followed by a new track receive a much smaller probability through the spatial affordance map than hypotheses that assume this to be an occlusion. The fact that the green graph in Figure 5 is below the ground truth indicates that the modified approach favors track occlusions slightly too much over deletion/creation pairs. The result however demonstrates clearly that the spatial affordance map enables a tracker to better hypothesize about the state of tracks, leading to a more accurate tracking behavior.

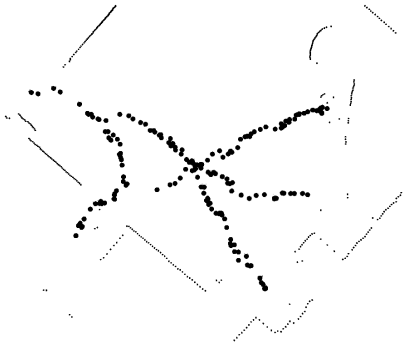

Fig. 4. Four (of 28) example tracks from experiment 1.

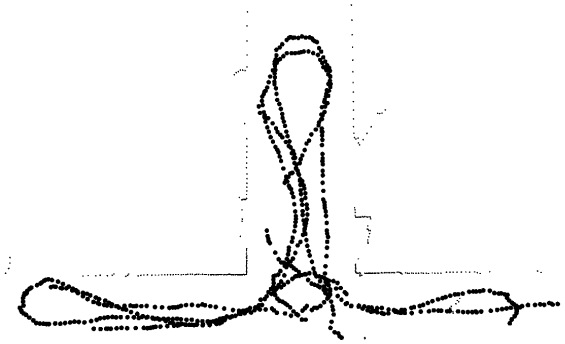

Fig. 6. Six (of 50) example tracks from experiment 2.

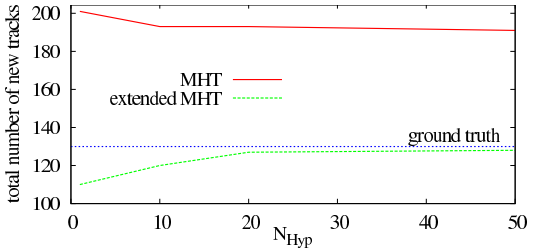

Fig. 5. The total number of tracks as a function of N_{Hyp}, the number of generated hypotheses. The tracking experiment had 4588 frames with a total of 130 people. The red line shows the MHT approach, the dotted green line the extended approach. The graph shows that replacing the fixed Poisson rates by the ones in the spatial affordance map improves the tracking accuracy significantly.

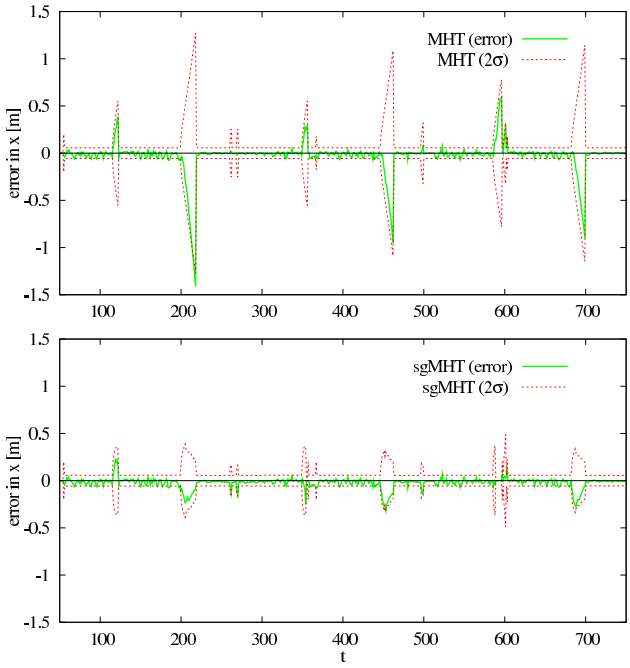

Fig. 7. Comparison between constant velocity motion model (top) and place-dependent motion model (bottom). Peaks correspond to occluded target maneuvers (turns around corners and U-turns). Fig. 3 shows the left turn of a person at step 217 of this experiment. While both approaches are largely consistent from a estimation point of view, the place-dependent model results in an overall smaller estimation error and smaller uncertainties. For eight manually inspected tracks, the constant velocity motion model lost a track three times while the new model had no track loss.

B. Place-Dependent Motion Model

In the second experiment, the constant velocity motion model is compared to our place-dependent motion model. A training set over 7443 frames with 50 person tracks in a office-like environment was recorded to learn the spatial affordance map (see Figure 6 and Figure 3). A test set with 1611 frames and eight people tracks was used to compare the two models. The data set was labeled by hand to determine both, the ground truth positions of people and the true data associations. In order to make the task more difficult, we defined areas in which target observations are ignored as if the person had been occluded by an object or another person. These areas were placed at hallway corners and U-turns where people typically maneuver. As the occlusion is simulated, the ground truth position of the targets is still available. As a measure of accuracy, the posterior position estimates of both approaches to the ground truth is calculated. The resulting estimation error in x is shown in Figure 7 (the error in y is similar).

The diagram shows much smaller estimation errors and 2σ bounds for the place-dependent motion model during target maneuvers. An important result is that the predicted covariances do not grow boundless during the occlusion events (peaks in the error plots). As illustrated in Figure 3, the shape of the covariance predictions follows the walkable area map at the very place of the target. Smaller covariances lead to lower levels of data association ambiguity, and thus, to decreased computational costs and more accurate probability distribution over pruned hypothesis trees.

For eight manually inspected tracks, the constant velocity motion model lost a track three times while the new model had no track loss. By tuning the entries of the process noise covariance Q, the constant velocity motion model can be made to avoid such losses, but this is clearly the wrong way to go as it brings along an even higher level of data association ambiguity.

VII. Conclusions

In this paper we presented an extended multi-hypothesis approach to laser-based people tracking that incorporates information on how people use space.

We proposed a non-homogeneous spatial Poisson process, called *spatial affordance map*, to represent the spatially varying distributions over track interpretation events of a MHT tracker and derive expressions for Bayesian learning of the map.

The spatial affordance map enabled us to relax and overcome the simplistic fixed Poisson rate assumption for new tracks and false alarms in the MHT approach. Using a learned spatio-temporal Poisson rate function, the system was able to compute refined probability distributions over hypotheses, resulting in a significantly more accurate tracking behavior in terms of steady track identities. The map further allowed us to derive a new, place-dependent model to predict target motion. The model showed superior performance in predicting maneuvering targets especially during lengthy occlusion events when compared to a constant velocity motion model.

In the future, we plan to extend the representation to a non-stationary Poisson process.

Acknowledgment

This work has partly been supported by the German Research Foundation (DFG) under contract number SFB/TR-8.

References

[1] B. Kluge, C. Köhler, and E. Prassler, "Fast and robust tracking of multiple moving objects with a laser range finder," in *Proc. of the Int. Conf. on Robotics & Automation (ICRA)*, 2001.

[2] A. Fod, A. Howard, and M. Mataríc, "Laser-based people tracking," in *Proc. of the Int. Conf. on Robotics & Automation (ICRA)*, 2002.

[3] M. Kleinhagenbrock, S. Lang, J. Fritsch, F. Lömker, G. Fink, and G. Sagerer, "Person tracking with a mobile robot based on multi-modal anchoring," in *IEEE International Workshop on Robot and Human Interactive Communication (ROMAN)*, Berlin, Germany, 2002.

[4] D. Schulz, W. Burgard, D. Fox, and A. Cremers, "People tracking with a mobile robot using sample-based joint probabilistic data association filters," *International Journal of Robotics Research (IJRR)*, vol. 22, no. 2, pp. 99–116, 2003.

[5] E. Topp and H. Christensen, "Tracking for following and passing persons," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, Alberta, Canada, 2005.

[6] J. Cui, H. Zha, H. Zhao, and R. Shibasaki, "Tracking multiple people using laser and vision," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, Alberta, Canada, 2005.

[7] M. Mucientes and W. Burgard, "Multiple hypothesis tracking of clusters of people," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, Beijing, China, 2006.

[8] G. Taylor and L. Kleeman, "A multiple hypothesis walking person tracker with switched dynamic model," in *Proc. of the Australasian Conf. on Robotics and Automation*, Canberra, Australia, 2004.

[9] J. Cui, H. Zha, H. Zhao, and R. Shibasaki, "Laser-based interacting people tracking using multi-level observations," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, Beijing, China, 2006.

[10] K. O. Arras, S. Grzonka, M. Luber, and W. Burgard, "Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities," in *Proc. of the Int. Conf. on Robotics & Automation (ICRA)*, 2008.

[11] D. B. Reid, "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, 1979.

[12] I. J. Cox and S. L. Hingorani, "An efficient implementation of reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, vol. 18, no. 2, pp. 138–150, 1996.

[13] Y. Bar-Shalom and X.-R. Li, *Multitarget-Multisensor Tracking: Principles and Techniques*. Storrs, USA: YBS Publishing, 1995.

[14] S. S. Blackman, "Multiple hypothesis tracking for multiple target tracking," *Aerospace and Electronic Systems Magazine, IEEE*, vol. 19, no. 1, pp. 5–18, 2004.

[15] A. Bruce and G. Gordon, "Better motion prediction for people-tracking," in *Proc. of the Int. Conf. on Robotics & Automation (ICRA)*, Barcelona, Spain, 2004.

[16] L. Liao, D. Fox, J. Hightower, H. Kautz, and D. Schulz, "Voronoi tracking: Location estimation using sparse and noisy sensor data," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2003.

[17] E. Mazor, A. Averbuch, Y. Bar-Shalom, and J. Dayan, "Interacting multiple model methods in target tracking: a survey," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 34, no. 1, pp. 103–123, Jan 1998.

[18] C. Kwok and D. Fox, "Map-based multiple model tracking of a moving object," in *RoboCup 2004: Robot Soccer World Cup VIII*, 2005, pp. 18–33.

[19] K. O. Arras, Óscar Martínez Mozos, and W. Burgard, "Using boosted features for the detection of people in 2d range data," in *Proc. of the Int. Conf. on Robotics & Automation (ICRA)*, Rome, Italy, 2007.

[20] K. Murty, "An algorithm for ranking all the assignments in order of increasing cost," *Operations Research*, vol. 16, 1968.