The International Journal of
Robotics Research
http://ijr.sagepub.com/

Place-dependent people tracking

Matthias Luber, Gian Diego Tipaldi and Kai O Arras

The International Journal of Robotics Research published online 17 January 2011
DOI: 10.1177/0278364910393538

The online version of this article can be found at:
http://ijr.sagepub.com/content/early/2011/01/08/0278364910393538

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
ijrr

Multimedia Archives

The International Journal of
Robotics Research
00(000) 1–14
© The Author(s) 2011
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0278364910393538
ijr.sagepub.com

Place-dependent people tracking

Matthias Luber, Gian Diego Tipaldi and Kai O Arras

Abstract

People detection and tracking are important in many situations where robots and humans work and live together. But unlike targets in traditional tracking problems, people typically move and act under the constraints of their environment. The probabilities and frequencies for when people appear, disappear, walk or stand are not uniform but vary over space making human behavior strongly place-dependent. In this paper we present a model that encodes spatial priors on human behavior and show how the model can be incorporated into a people-tracking system. We learn a non-homogeneous spatial Poisson process that improves data association in a multi-hypothesis target tracker through more informed probability distributions over hypotheses. We further present a place-dependent motion model whose predictions follow the space-usage patterns that people take and which are described by the learned spatial Poisson process. Large-scale experiments in different indoor and outdoor environments using laser range data, demonstrate how both extensions lead to more accurate tracking behavior in terms of data-association errors and number of track losses. The extended tracker is also slightly more efficient than the baseline approach. The system runs in real-time on a typical desktop computer.

Keywords

Cognitive robotics, human-centered and life-like robotics, learning and adaptive systems, sensing and perception computer vision, social human–robot interaction, visual tracking

1. Introduction

As robots start to enter domains in which they interact and cooperate closely with humans, people tracking is becoming a key technology for areas such as human–robot interaction, human activity understanding or intelligent cars. In contrast to most airborne and waterborne targets, people typically move and act under environmental and social constraints. These constraints vary over time and space, making possible motion and action patterns strongly place-dependent. Examples include walls through which people cannot walk or a cooking stove, which limits the activity of cooking to a specific place, etc.

In this paper we learn and represent human spatial behavior for the purpose of people tracking. 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 people in the world, and to make place-dependent predictions of human motion that reflect how people are actually using space. We propose a non-homogeneous spatial Poisson process to represent the spatially varying distribution over relevant human-activity events. This representation, called a *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.

In most related work on laser-based people tracking (Kluge et al. 2001; Fod et al. 2002; Kleinhagenbrock et al. 2002; Schulz et al. 2003; Cui et al. 2005; Topp and Christensen 2005; Mucientes and Burgard 2006), a person is represented as a single state that encodes torso position and velocity. People are extracted from range data as single blobs or found by merging nearby point clusters that correspond to legs. People tracking has also been addressed as a leg-tracking problem (Taylor and Kleeman 2004; Cui et al. 2006; Arras et al. 2008) where people are represented by the states of both legs, either in a single augmented state

Social Robotics Laboratory, University of Freiburg, Department of Computer Science, Germany

Corresponding author:
Matthias Luber, Social Robotics Laboratory, University of Freiburg, Department of Computer Science, Georges-Koehler-Allee 106, 79110 Freiburg, Germany.
Email: luber@informatik.uni-freiburg.de

(Cui et al. 2006) or as a high-level track to which two low-level leg tracks are associated (Taylor and Kleeman 2004; Arras et al. 2008).

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 by Fod et al. (2002), Kluge et al. (2001) and Kleinhagenbrock et al. (2002). A sample-based Joint Probabilistic Data Association Filter (JPDAF) has been presented by Schulz et al. (2003) and adopted by Topp and Christensen (2005). And in Khan et al. (2006) a Markov chain Monte Carlo (MCMC)-based auxiliary variable particle filter is proposed. The Multi-Hypothesis Tracking (MHT) approach according to Reid (1979) and Cox and Hingorani (1996) has been used by Taylor and Kleeman (2004), Mucientes and Burgard (2006) and Arras et al. (2008). What makes MHT an attractive choice is that it belongs to the most general type of 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, JPDAF or PMHT by Streit and Luginbuhl (1995) are suboptimal in nature as they simplify the problem (Bar-Shalom and Li 1995; Blackman 2004). For these reasons, MHT has become a widely accepted tool in the target-tracking community, as pointed out by Blackman (2004), especially for problems with a low to medium number of targets.

For people tracking, however, the MHT approach relies on statistical assumptions that are overly simplified and do not account for place-dependent target behavior. The approach assumes that new tracks and false alarms are uniformly distributed in the sensor's field of view with fixed Poisson rates. While this might be acceptable in the settings for which the approach was originally developed (using, e.g., radar or underwater sonar), it does not account for the non-random usage of an environment by people. Human subjects appear, disappear, walk or stand at specific locations. False alarms are also more likely to arise in areas with cluttered backgrounds rather than in open spaces. A simple form of place-dependency has been realized in Breitenstein et al. (2009), a visual surveillance scenario with a static camera, where a frame around the border of the image was manually positioned to indicate the area where new tracks may appear. In the center of the image, no new tracks are assumed to arrive. In this paper, we extend prior work by incorporating learned distributions over track-interpretation events in order to support data association and show how a non-homogeneous spatial Poisson process can be used to seamlessly extend the MHT approach for this purpose.

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 (2004) and Liao et al. (2003).

Bruce and Gordon (2004) learn 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. (2003) extract a Voronoi graph from a map of the environment and represent the states of people as the edges of that graph. This allows them to predict the motion of people along 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's behavior are a widely accepted technique to deal with maneuvering targets, of note is the Interacting Multiple Model (IMM) algorithm (for a survey, see Mazor et al. 1998). Different target-motion models have also been studied by Kwok and Fox (2005). 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.

Regarding motion models, our approach extends prior work in two aspects: learning and place-dependency. In contrast to Liao et al. (2003) and Kwok and Fox (2005) 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 Liao et al. (2003), the positions of people are 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 into why people move on a Voronoi set, particularly in open spaces whose topology is not well defined. Our approach, by contrast, tracks the actual position of people and predicts their motion according to metric, place-dependent models. Contrary to Bruce and Gordon (2004) where motion prediction is made along paths that a planner creates 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 a brief overview of the people tracker that will later be extended, introducing the theory of the spatial affordance map and expressions for learning its parameters. Section 3 describes how the map is used to improve data association from refined probability distributions over hypotheses, while Section 4 presents the theory of the place-dependent motion model. Section 5 presents the experimental results followed by the conclusions in Section 6.

2. Spatial affordance map

We pose the problem of learning a spatial model of human behavior as a parameter-estimation problem of a non-homogeneous spatial Poisson process. The resulting model, called a *spatial affordance map*, is a global long-term representation of human-activity events in an environment. The name lends itself to the concept of affordances as we consider the possible sets of human actions and motions as a result of environmental constraints. An affordance is a resource or support that an object (the environment) offers an (human) agent for action. This section describes the theory and how learning in the spatial affordance map is implemented.

A Poisson distribution is a discrete distribution of 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 unit. As we are interested 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. Let $N(t)$ be a discrete random variable, which represents the number of events occurring up to time t with rate λ. Then, $N(t)$ 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 times a and b is

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

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 $\lambda(x, t)$ with $x \in X$ 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

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

If this generalized rate function is a separable function of time and space, we have

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

for some function $f(x)$ for which we can demand

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

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

Learning the spatio-temporal distribution of events in an environment is equivalent to learning the generalized rate function $\lambda(x, t)$. 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 $\lambda(x, t)$ using a grid approximation,

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

with $1_{ij}(x)$ being the indicator function defined as $1_{ij}(x) = 1$ if $x \in \text{cell}_{ij}$ and $1_{ij}(x) = 0$ if $x \notin \text{cell}_{ij}$. This type of approximation is not imperative and there is no loss of generality. Other space tessellation techniques such as graphs, quadtrees or arbitrary regions with homogeneous Poisson rates can equally be used. The subdivision of space into regions of fixed Poisson rates has the property that the preferable decomposition in Equation (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 each type of event. For the purpose of this paper, the map has three layers, one for new tracks, one for matched tracks, and one for false alarms. The first layer represents the distribution and rates of people appearing in the environment. The second layer can be considered as 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.

2.1. 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,

$$\text{Gamma}(\lambda; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \qquad \text{for } \lambda > 0. \tag{8}$$

Then, learning the rate parameter λ consists of 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 $P(\lambda_i|K_{1..i}) \sim P(k_i|\lambda_{i-1})\, P(\lambda_{i-1})$ 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 Equation (6) the likelihood. Then by substitution, it can be shown that the update rules for the parameters are $\alpha_i = \alpha_{i-1} + k_i$ and $\beta_i = \beta_{i-1} + 1$. The posterior mean of the rate parameter in a single cell is finally obtained as the expected value of the Gamma,

$$\widehat{\lambda}_{\text{Bayesian}} = \mathbb{E}[\lambda] = \frac{\alpha}{\beta} = \frac{\#\text{positive events} + 1}{\#\text{observations} + 1}\,. \tag{9}$$

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 proper initialization of 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 $\lambda(x, t)$. 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 Equation (4) holds.[1] In this case, the probability distribution function (pdf), $f(x)$, is given by

$$f(x) = \frac{\lambda(x, t)}{\lambda(t)} \tag{10}$$

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

$$\lambda(t) = \int_X \lambda(x, t)\, \mathrm{d}x. \tag{11}$$

In our grid, those quantities are computed as

$$f(x) = \frac{\sum_{(i,j)\in X} \lambda_{ij} 1_{ij}(x)}{\sum_{(i,j)\in X} \lambda_{ij}}. \tag{12}$$

If there are several layers in the map, each layer contains the distribution $f(x)$ 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 1 shows two layers of the spatial affordance map of our laboratory, learned during the first experiment described in Section 5. The picture on the left shows the space-usage distribution of the environment. The modes in this distribution correspond to often used places and correspond to goal locations in that room (three desks and a couch). On the right, the distribution of new tracks is depicted. Peaks indicate locations where people appear (doors). The reason for the small peaks at locations other than the doors is that when subjects interact with objects (sit on a chair, lie on the couch), the tracker loses them. When they re-enter the space, they are detected as new tracks.

3. Data association with spatial-target priors

Many tracking approaches rely on rather simple models for new-track and false-alarm events and ignore important information that is available from the environment. For example, the MHT approach 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's field of view V. While this is a valid assumption for a radar aimed upwards into the sky, it does not account for the place-dependent character of human behavior. The way that people move is often due to environmental constraints that can be learned. Indoors, for instance, doors or convex corners are typical places where people appear. This place-dependency is seen by a detector. Regions of clutter and complex background produce false alarms more often than in open space, making a spatially uniform model a poor approximation.

The spatial affordance map correctly holds the necessary information. In this paper we extend the original MHT of Reid (1979) and Cox and Hingorani (1996) with spatial priors and show that the map allows for a seamless integration into the MHT framework. In particular, we replace the temporal fixed-rate models for new tracks and false alarms by the learned Poisson rates for arrival events of people and false detections, and the spatial uniform probability with the learned location statistics.

We will first give a short outline of the regular MHT approach. In summary, the MHT 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 tracks, and tracks as matched, occluded or obsolete. A hypothesis Ω_i^t is one possible set of assignments and interpretations at time 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 people detected in the laser data. For detection, we use a learned classifier based on a collection of boosted features, see Arras et al. (2007). Let $\psi_i(t)$ denote a set of assignments that associates predicted tracks with 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

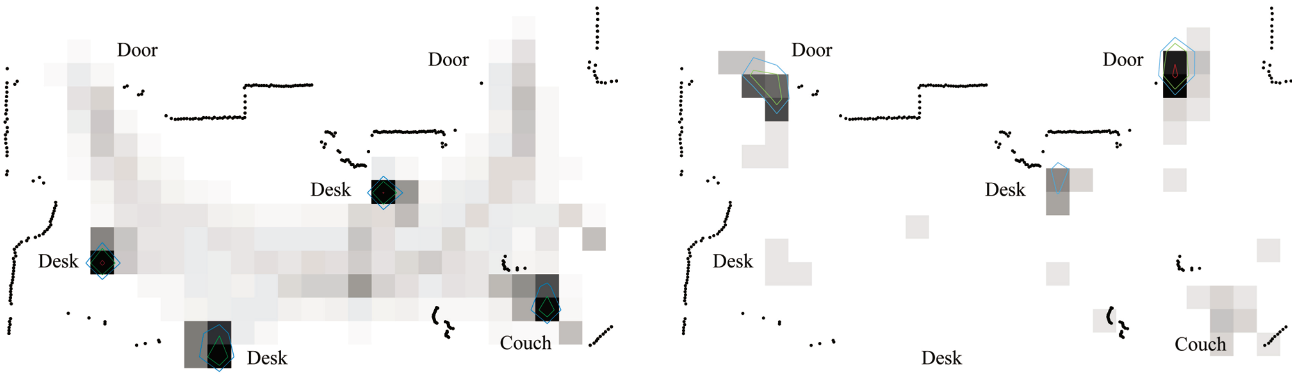

Fig. 1. Spatial affordance map of the laboratory in experiment 1. The probability distribution of matched track events is shown on the left, the distribution of new track events is shown on the right. The marked locations in each distribution (extracted with a peak finder and visualized by contours of equal probability) have different meanings. On the left they correspond to places that are often visited by people (three desks and a couch), while the maxima of the new-track distribution (right) denote locations where people appear in the sensor's field of view (two doors, the couch and a desk).

the parent. This creates a hypothesis tree that grows exponentially. 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,

$$\begin{aligned} p\left(\Omega_i^t \mid Z^t\right) = \eta \cdot p\left(Z(t) \mid \psi_i(t), \Omega_{p(i)}^{t-1}\right) \\ p\left(\psi_i(t) \mid \Omega_{p(i)}^{t-1}, Z^{t-1}\right) \cdot p\left(\Omega_{p(i)}^{t-1} \mid Z^{t-1}\right). \end{aligned} \quad (13)$$

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 with a track t_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's field of view, with probability V^{-1}. Thus

$$p\left(Z(t) \mid \psi_i(t), \Omega_{p(i)}^{t-1}\right) = V^{-(N_F+N_N)} \cdot \prod_{i=1}^{m_t} \mathcal{N}(z_i(t))^{\delta_i} \quad (14)$$

with N_F and N_N being the number of measurements labeled as false alarms and new tracks, respectively. Here δ_i is an indicator variable being 1 if measurement i is associated with 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

$$p_N(x) = \frac{\lambda_N(x,t)}{\lambda_N(t)} = \frac{\lambda_N(x,t)}{\int_V \lambda_N(x,t)\;\mathrm{d}x} \quad (15)$$

where $\lambda_N(x,t)$ is the learned Poisson rate of new tracks in the map and x the position of measurement $z_i'(t)$ transformed into global coordinates. Given our grid, Equation (15) becomes

$$p_N(x) = \frac{\lambda_N(z_i'(t), t)}{\sum_{(i,j)\in V} \lambda_{ij,N}}. \quad (16)$$

The probability of false alarms $p_F(x)$ is calculated in the same way using the learned Poisson rate of false alarms $\lambda_F(x,t)$ in the map. Although the theory presented so far is general, in this paper we assume the behavior of people when appearing and the false-positive statistics of the detector are time-invariant, and, therefore, the Poisson process is only non-homogeneous over space. The rate parameters $\lambda_N(x,t)$ and $\lambda_F(x,t)$ then become $\lambda_N(x)$ and $\lambda_F(x)$, respectively.

As presented by Arras et al. (2008), the expression for the assignment set probability in the MHT can be shown to be

$$\begin{aligned} p\left(\psi_i(t) \mid \Omega_{p(i)}^{t-1}, Z^{t-1}\right) = \eta' \cdot p_M^{N_M} \cdot p_O^{N_O} \cdot p_D^{N_D} \cdot \\ \lambda_N^{N_N} \cdot \lambda_F^{N_F} \cdot V^{(N_F+N_N)} \end{aligned} \quad (17)$$

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 and $p_M + p_O + p_D = 1$. 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 Equation (13) means that, as in the original approach, many terms cancel out leading to an easy-to-implement expression for the hypothesis probability

$$p\left(\Omega_i^t \mid Z^t\right) = \eta'' \cdot p_M^{N_M} \cdot p_O^{N_O} \cdot p_D^{N_D} \cdot \prod_{i=1}^{m_t} \Big[\mathcal{N}(z_i(t))^{\delta_i} \cdot \lambda_N(z_i'(t))^{\kappa_i} \cdot \lambda_F(z_i'(t))^{\phi_i} \Big] \cdot p\left(\Omega_{p(i)}^{t-1} \mid Z^{t-1}\right) \qquad (18)$$

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

The insight into this extension of MHT is that we replace fixed parameters by spatial priors on human behavior in the form of learned spatial rate functions. As we will show, this domain knowledge will lead to refined probability distributions over hypotheses and helps the tracker to better interpret measurements and tracks. This extension comes at no additional runtime costs.

4. Place-dependent motion model

People are highly dynamic targets to track. They can abruptly stop, turn back, left or right, make a sideways step or accelerate. However, human motion is not random but follows place-dependent patterns typically formed by the environment: people turn around convex corners, maneuver around obstacles, stop in front of doors and do not go through walls. The Brownian model, the constant-velocity and even higher-order motion models are clearly unable to capture the complexity of these movements. In addition to this, people often undergo lengthy occlusion events during interaction with each other or with the environment. In this section we propose a place-dependent motion model for short-term predictions of maneuvering targets. It relies on learned human-motion priors in order to account for this complexity.

Formally, this means that the motion model $p(x_t|x_{t-1}, m)$ becomes conditioned on both the previous track state x_{t-1} and the walkable-area map m obtained by clipping the space-usage probability defined in Equation (12) at a given probability. It describes a general density that follows the shape and topology of the environment, poorly described by a parametric distribution such as a Gaussian. We, therefore, follow a sampling approach and represent our target distribution with a set of weighted samples

$$p(x_t|x_{t-1}, m) \simeq \sum_i w_t^{(i)} \delta_{x_t^{(i)}}(x_t) \qquad (19)$$

where $\delta_{x_t^{(i)}}(x_t)$ is the impulse function centered on $x_t^{(i)}$.

Sampling directly from the distribution $p(x_t|x_{t-1}, m)$ is intractable in practice, which is why we take a Monte Carlo approach, in which samples are first drawn from a proposal distribution π and then evaluated according to the mismatch between the target distribution τ and the proposal distribution. In our case, the distribution is approximated by the following factorization

$$p(x_t|x_{t-1}, m) \simeq p(x_t|x_{t-1}) \cdot p(x_t|m) \qquad (20)$$

and we adopt the natural choice by using a motion model $p(x_t^{(i)}|x_{t-1})$ as our proposal distribution and evaluate the samples according to

$$w_t^{(i)} = \frac{p(x_t|x_{t-1}, m)}{p(x_t^{(i)}|x_{t-1})} = p(x_t^{(i)}|m). \qquad (21)$$

In other words, samples are first distributed into the state space following the motion model $p(x_t^{(i)}|x_{t-1})$ and then weighted according to the map m.

For $p(x_t^{(i)}|x_{t-1})$, we take the curvilinear model of Best and Norton (1997). This motion model is simple, yet is one of the most sophisticated target-maneuver models in 2D as pointed out by Rong Li and Jilkov (2003). It accounts for both (cross-track) normal and (along-track) tangential target accelerations. As illustrated in Figure 2, constant-velocity and constant-turn motion follow as special cases. Let $x_t^{(i)} = (x_t \; y_t \; \dot{x}_t \; \dot{y}_t)^{\mathrm{T}}$ be the state of particle i at discrete time t, $a_t = (a_t \; a_n)^{\mathrm{T}}$ the vector of tangential and normal accelerations, and A_{cv} the transition matrix of the constant-velocity model, then the particle states evolve according to

$$x_{t+1}^{(i)} = A_{cv} x_t^{(i)} + G_t(a_t^{(i)} + q_t) \qquad (22)$$

with q_t being zero-mean Gaussian noise with covariance matrix Q_t. Further details of the 4 × 2 forcing matrix G_t can be found in Best and Norton (1997).

At each discrete time t and for each track, samples are drawn from the posterior state estimate $(x_t, \; \Sigma_t)$ and sent into different directions by randomizing the accelerations a_t by a noise with covariance $Q_t = \mathrm{diag}[\sigma_{a_t}^2, \; \sigma_{a_n}^2]$ (see Figure 2, right). When an occlusion event occurs, the particles will evolve through Equation (22) and are weighted and resampled according to the strategy described below.

Even a sophisticated motion model can strongly differ from our target distribution, especially at places where the walkable area is highly constrained by the environment. This means that many samples fall into low probability regions leading to the known problem of particle depletion. For this reason, we follow the auxiliary particle filter approach of Pitt and Shephard (1999), which was developed for such mismatch situations. In a nutshell, the auxiliary particle filter computes an improved proposal derived from an approximated observation likelihood. In our case, this feature can be extended to a look-ahead ability since the map m delivers the observation likelihood that can be probed at locations computed by forward-simulating the motion model.

Assume that we have a set of samples at discrete time $t-1$, which represents our target distribution. The distribution at time t is then

$$p(x_t|x_{t-1}, m) \simeq p(x_t|m) \sum_i p\left(x_t|x_{t-1}^{(i)}\right) w_{t-1}^{(i)}. \qquad (23)$$

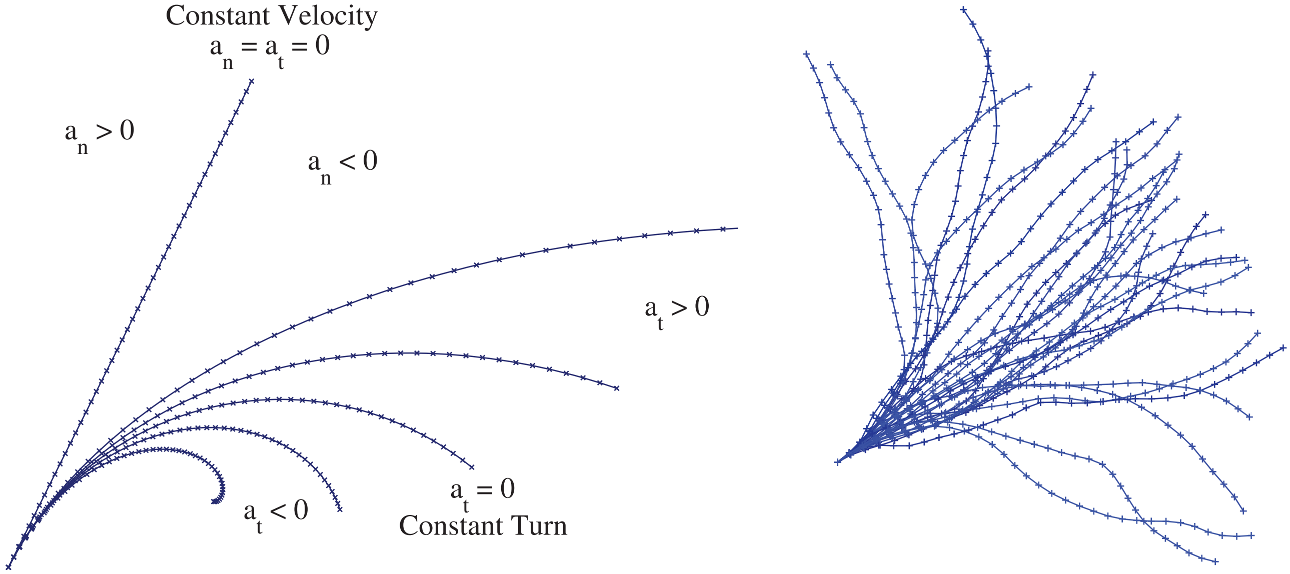

Fig. 2. Curvilinear-motion model according to Best and Norton (1997). *Left:* The model accounts for both (cross-track) normal acceleration a_n and (along-track) tangential acceleration a_t. *Right:* Sample particles over 30 steps at $\Delta t = 0.1$ s, subject to white zero-mean Gaussian noise with $\sigma_{a_t} = 0.1$ m s^{-2} and $\sigma_{a_n} = 1$ m s^{-2}.

To avoid depletion, and following Pitt and Shephard (1999), we sample from the higher dimensional distribution $p(x_t, k|x_{t-1}, m)$, where the auxiliary variable k denotes the index of the sample at time $t-1$ in the mixture defined above, to give

$$p(x_t, k|x_{t-1}, m) \simeq p(x_t|m)\, p\left(x_t|x_{t-1}^{(k)}\right) w_{t-1}^{(i)}, \qquad (24)$$

and ignoring the sampled index, we obtain a sample from the original target density. Equation (24) can be approximated by

$$g(x_t, k|x_{t-1}, m) \simeq p\left(\mu_t^{(k)}|m\right) p\left(x_t|x_{t-1}^{(k)}\right) w_{t-1}^{(i)}, \qquad (25)$$

where $\mu_t^{(k)}$ is the mean, the mode, a draw, or some other value associated with the density of the $p(x_t|x_{t-1}^{(k)})$ used to evaluate the goodness of the parent sample $x_{t-1}^{(k)}$. The approximated density is designed such that we can sample from $g(x_t, k|x_{t-1}, m)$ by first sampling the index according to the pseudo-weight $\lambda_k \propto g(k|x_{t-1}, m)$ and then sampling from the corresponding motion model $p(x_t|x_{t-1}^{(k)})$, where

$$g(k|x_{t-1}, m) = \int p\left(\mu_t^{(k)}|m\right) p\left(x_t|x_{t-1}^{(k)}\right) w_{t-1}^{(i)}\, \mathrm{d}x_t \qquad (26)$$

$$= p\left(\mu_t^{(k)}|m\right) w_{t-1}^{(i)}. \qquad (27)$$

The weights of these new samples are finally computed by

$$w_t = \frac{p\left(x_t|m\right) p\left(x_t|x_{t-1}^{(k)}\right) w_{t-1}^{(i)}}{p\left(\mu_t^{(k)}|m\right) p\left(x_t|x_{t-1}^{(k)}\right) w_{t-1}^{(i)}} = \frac{p\left(x_t|m\right)}{p\left(\mu_t^{(k)}|m\right)}. \qquad (28)$$

In our case, we use the above mentioned curvilinear-motion model to compute a *look-ahead particle* as the future estimate $\mu_t^{(k)}$. This is done by propagating the kth sample l steps into the future, that is, the sample is forward-simulated via the motion model over a time interval $l\,\Delta t$. The value that is finally taken for $p(\mu_t|m)$ is then the value of $p(x_t|m)$ evaluated at the position of the look-ahead particle.

Once we have obtained the new motion model in the form of a weighted sample set, we need to integrate it into the MHT framework. Since MHT relies on the Kalman filter for tracking, the first two moments are computed as

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

$$\hat{\Sigma} = \frac{1}{1 - \sum_i (w^{(i)})^2} \sum_i w^{(i)} \left(\hat{\mu} - x_t^{(i)}\right)\left(\hat{\mu} - x_t^{(i)}\right)^{\mathrm{T}} \qquad (29)$$

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. Sample situations that illustrate the place-dependent motion model are shown in Figure 3.

5. Experiments

For the experiments, we collected four data sets, two in indoor and two in outdoor environments. The data sets are from a laboratory (Figure 4), an office building (Figure 8), the main station of Freiburg and a busy pedestrian zone in downtown Freiburg (Figures 5 and 6). The sensor was a SICK LMS 291 laser scanner with an angular resolution of 0.5° mounted at a height of 0.85 m with an acquisition rate of 12 Hz.

The spatial affordance maps were trained with the baseline MHT tracker of Cox and Hingorani (1996) with a detection probability of $p_{\mathrm{det}} = 0.999$, a termination likelihood $\lambda_{\mathrm{del}} = 20$, and 300 hypotheses. The parameters of

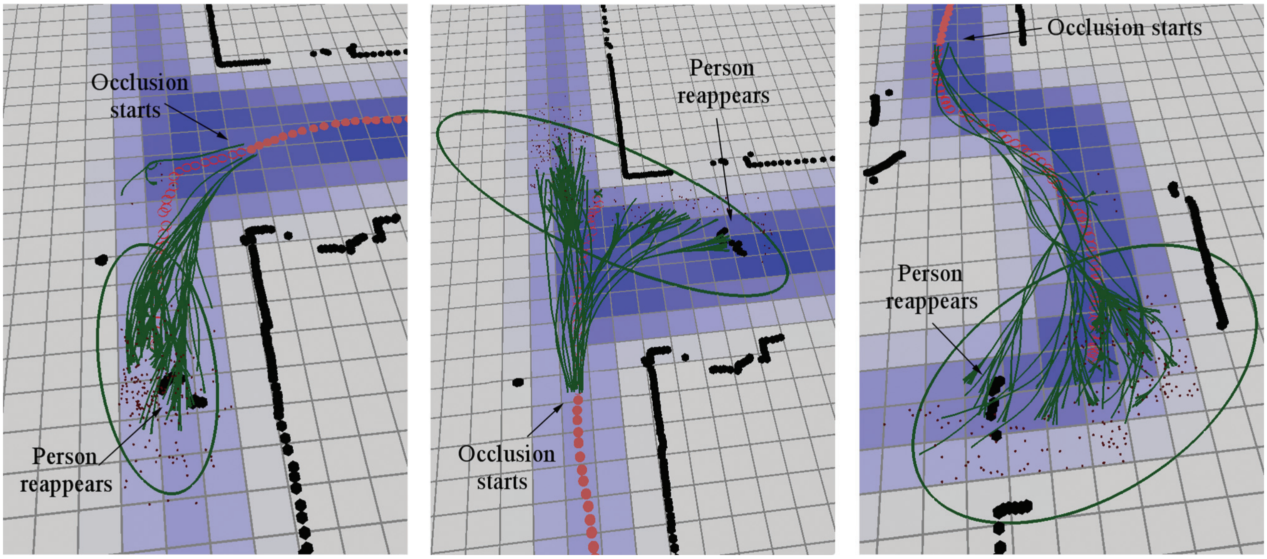

Fig. 3. Place-dependent motion model in three sample situations. The figures show a maneuvering target that reappears after a very long occlusion event. The background grid contains the learned space-usage probabilities of the spatial affordance map, thick black dots are laser measurements, small dots are the look-ahead particles, and the green ellipses illustrate the weighted 99% sample covariance from the particles. The model is able to predict the targets 'around the corner' and along the high-probability ridges in the map, yielding correct motion predictions for these types of situation.

the tracker were learned from training data with 95 labeled tracks over 28,242 frames. All data associations including occlusions were hand-labeled. This led to a fixed Poisson rate for new tracks $\lambda_N = 0.0002$ and a fixed Poisson rate of false alarms $\lambda_F = 0.0041$. The rates were estimated using the Bayesian approach in Equation (9). Care was taken to ensure that the estimates of the expected number of events were normalized with the sensor's field of view V. The grid cells of the map were chosen to be 30 cm in size. After the learning phase, the map was assumed to be fixed. For a pruning strategy, we employed N-scan-back logic with a tree depth of 30 and limited the maximum number of hypotheses to N_{Hyp} using the multi-parent variant of the algorithm proposed by Murty (1968).

The parameters of the place-dependent motion model were set to 300 samples, $\sigma^2_{a_t} = 0.1$ and $\sigma^2_{a_n} = 0.8$ for the noise for the tangential and normal accelerations, respectively, and $l = 5$ as a look-ahead factor to compute the pseudo-weights λ_k.

The data sets including annotations are available on the Webpage of the authors.

5.1. MHT data association with spatial target priors

In the first experiment, the original MHT approach is compared to the tracker using the spatial affordance map on the laboratory data set of over 38,994 frames and with a total of 134 people entering and leaving the sensor's field of view. As mentioned, the data association ground truth of the 134 tracks was determined manually.

To compare the impact of the presented models on tracking performance, we first tested the individual models against the baseline tracker and then evaluated the combination of both models. The accuracy of the resulting strategies was measured using the CLEAR MOT metric proposed by Bernardin and Stiefelhagen (2008). The metric has three numbers with respect to the ground truth that are incremented at each frame: misses (missing tracks that should exist at a ground truth position), false positives (tracks that should not exist), and mismatches (track identifier switches). The latter value quantifies the ability to deal with occlusion events. From these numbers, two values are determined: MOTP (average metric distance between estimated targets and ground truth) and MOTA (the average number of occurrences of the correct tracking output with respect to the ground truth). We ignore MOTP as it is based on a metric ground truth of target positions, which is unavailable in our data. In order to show the evolution of the error as a function of N_{Hyp}, which is proportional to the computational effort, N_{Hyp} is varied from 10 to 500.

The results show a significant improvement for the extended MHT with spatial priors over the regular approach, especially for the number of mismatches (see Figure 4). For $N_{Hyp} = 500$ the tracker made 107 fewer id switches (192 vs. 85), the number of false positives decreased from 4,930 to 2,624, and the number of misses from 383 to 272. The accuracy (MOTA) increased from 85% to 92%. The place-dependent new-track and false-alarm models even applied in isolation (blue and green lines) gave a performance increase over the baseline MHT.

The insights into these improvements are as follows. As can be seen in Figure 1 right, few new-track events were

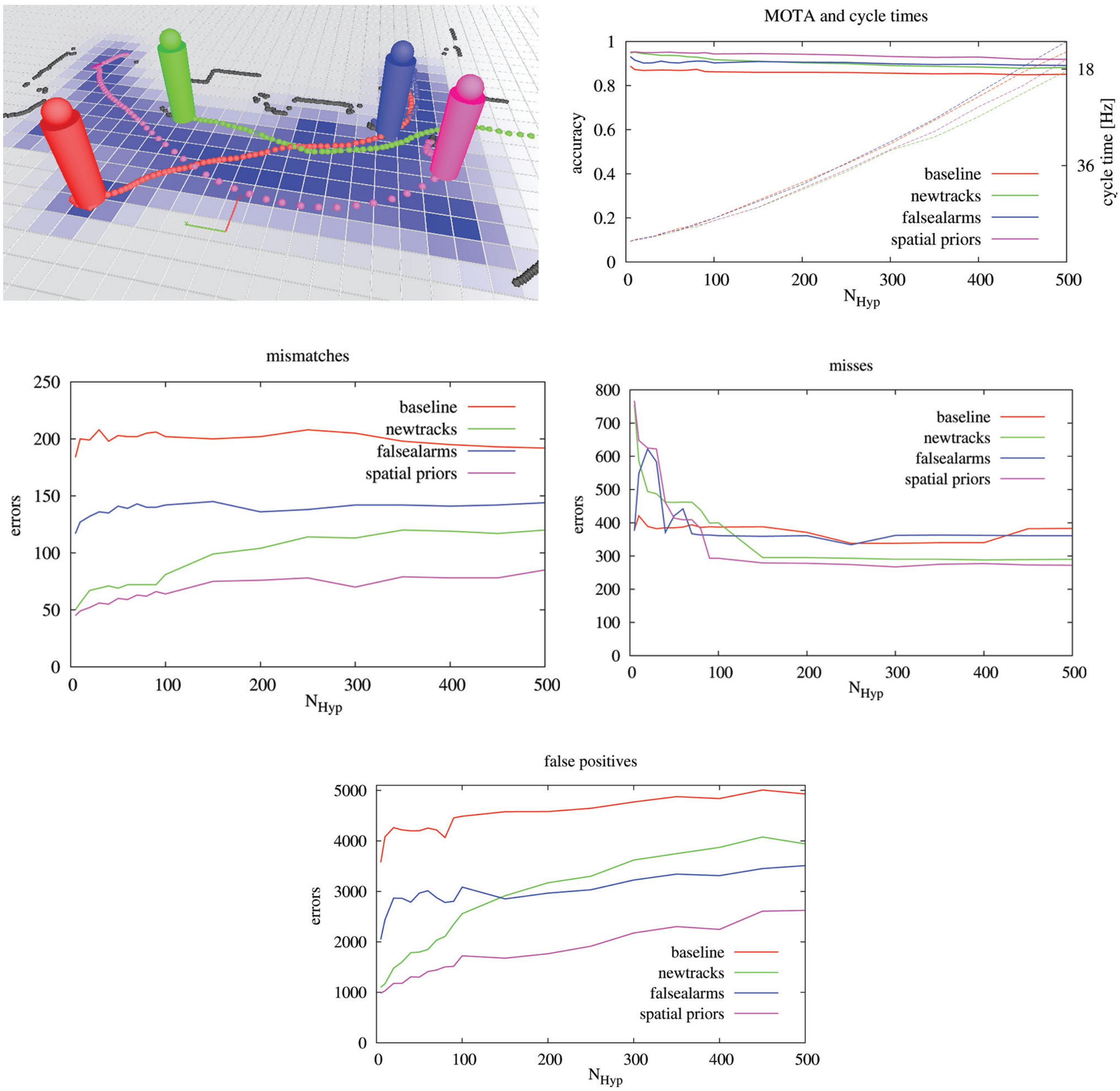

Fig. 4. Four of 134 sample tracks from the laboratory data set (top left). Accuracy of the different tracking approaches (MOTA, top right), total number of mismatches, misses and false positives as a function of N_{Hyp} (bottom, from left to right). The solid red lines show the results of the baseline MHT with fixed Poisson rates for new tracks and false alarms. The green and blue lines are for the extended approach using the spatial priors for new tracks and false alarms, respectively. The results for the combined approach are denoted by the magenta lines. The tracker cycle times are the lower, dotted lines in the top right diagram. The graphs show that when replacing the fixed Poisson rates by the learned, place-dependent ones, the tracker makes significantly fewer errors at slightly faster cycle times.

observed in the center of the room. If, for instance, a track occlusion occurs at such a place (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 improvement in the false-positive error is explained by fewer incorrect track creations in regions of clutter. This is due to both lower new-track probabilities at such places and higher false-alarm rates in regions of clutter. The combined approach benefits from both aspects and further reduces this type of error. Fewer misses are due to earlier track creations. The modes in the new-track distribution, especially around doors, allow the system to initialize tracks faster than with a fixed rate. Short sequences of observations are also tracked more accurately causing fewer errors of this type.

An additional data set with 15 people was collected to investigate whether the model is overfitted and generalizes

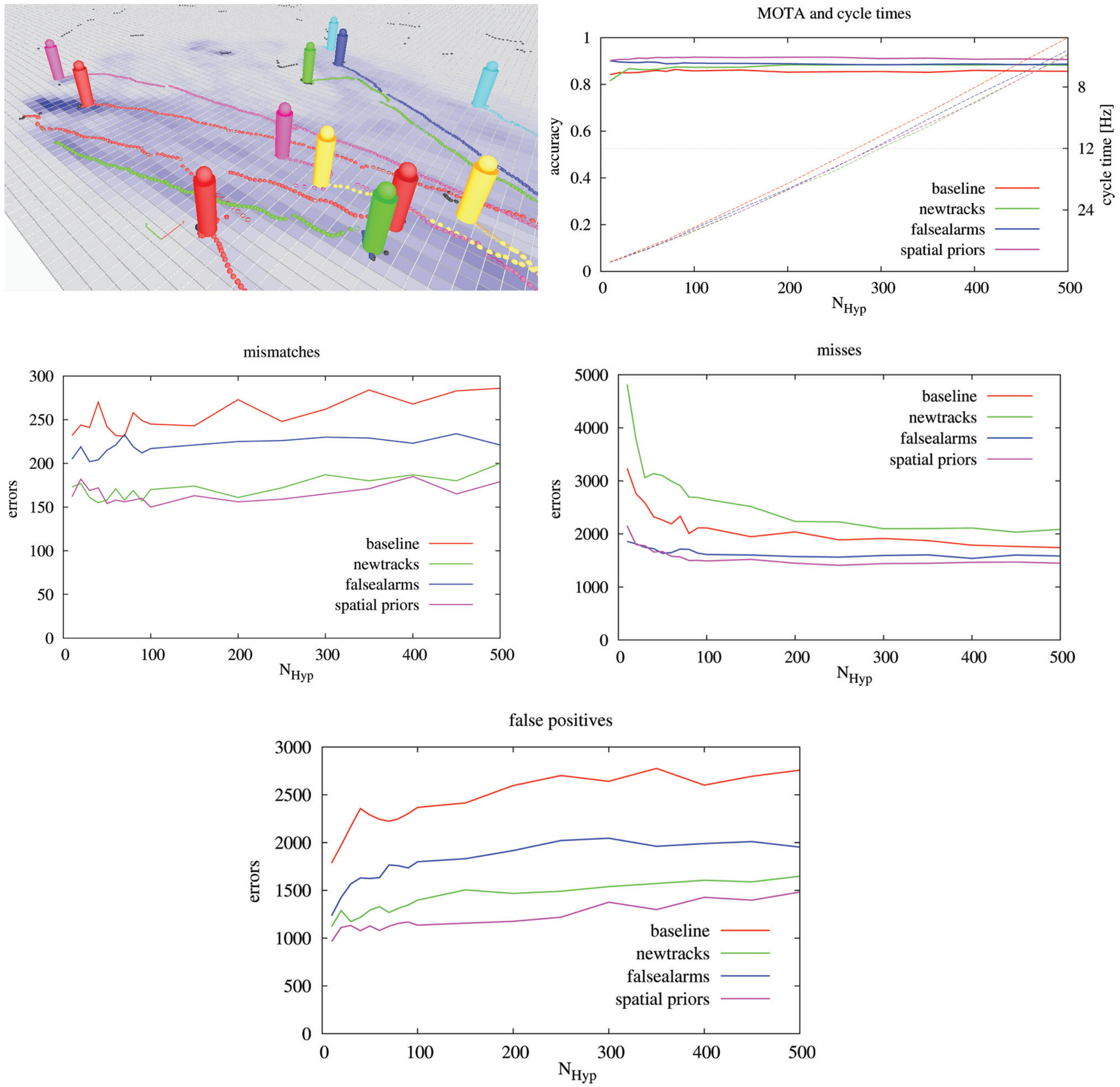

Fig. 5. From the experiment in an underground hall of Freiburg's main station, 12 of 160 sample tracks (top left). Accuracy of the different tracking approaches (MOTA, top right) and total number of mismatches, misses and false positives as a function of N_{Hyp} (bottom, from left to right). The red line shows the baseline MHT with fixed Poisson rates, the green and blue lines are for the system extended by the new-track and false-alarm models, the magenta line gives the combined approach. As in the indoor experiment, the diagrams show that the combined approach significantly reduces the number of mismatches, false positives and misses.

poorly when people behave in an unusual way. In this experiment, subjects entered the sensor's field of view through entry points that were not used previously (in between the couch and the desk at the bottom in Figure 1) or appeared in the center of the room by jumping off tables. Manual inspection of the resulting trees (using the graphviz-lib for visualization) revealed that the 15 people were tracked correctly. The difference with the approach for fixed Poisson rates is that, after track creation, the best hypothesis is not the true one during the first few (less than five) iterations. However, the incorrect hypotheses that successively postulate that the subjects are false alarms become very unlikely, causing the algorithm to backtrack to the true hypothesis within milliseconds.

To demonstrate the scalability of our extensions, we evaluated them in two unscripted large-scale outdoor settings. The first data set was collected in an underground hall in Freiburg's main station and the second one in a pedestrian zone in the city center of Freiburg during a regular workday. The data sets consist of 33,204 frames during

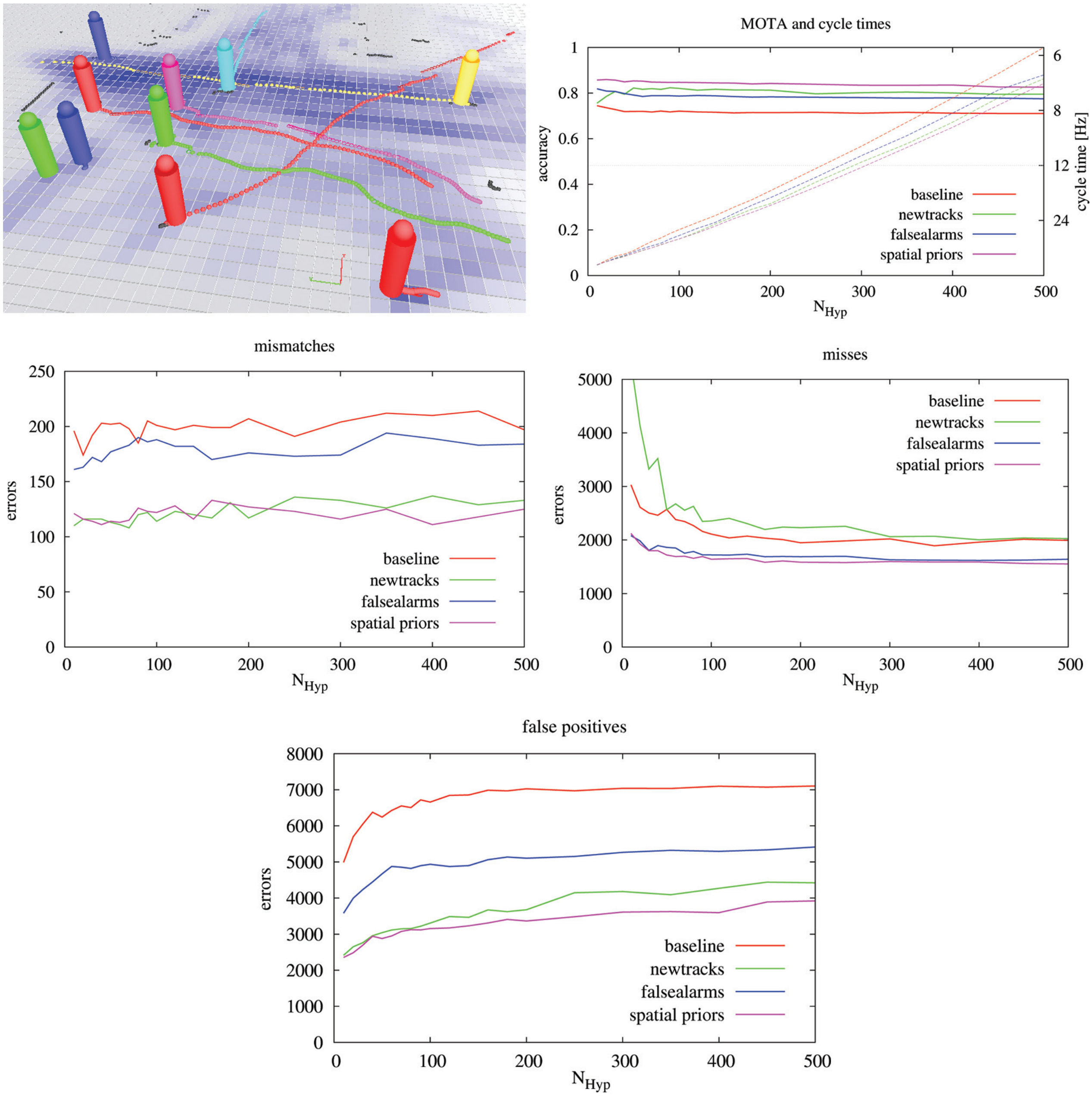

Fig. 6. From the experiment in the city center of Freiburg, 10 of 168 sample tracks (top left). Accuracy of the different tracking approaches (MOTA, top right) and total number of mismatches, misses and false positives as a function of N_{Hyp} (bottom, from left to right). The red, green, blue and magenta lines denote the baseline, the place-dependent false-alarm and new-track models and the combined approach, respectively. Again, the diagrams show that the place-dependent models significantly improve tracking performance.

15 minutes and 55,475 frames during 25 minutes, respectively. To determine the data association ground truth, 6,000 frames with 160 persons and 10,000 frames with 168 persons were again labeled by hand. The data sets contain up to 19 simultaneously visible targets with very frequent occlusions from other individuals or obstacles in the environment.

The results of the first outdoor experiment at Freiburg's main station show that the extended MHT with spatial priors yields similar improvements to the indoor data set (see Figure 5). At $N_{Hyp} = 500$, the accuracy (MOTA) increased from 85% to 91%. A detailed analysis of the CLEAR MOT metric shows that the number of mismatches dropped from 286 to 179, the number of false positives from 2,758 to 1,483 and the number of misses decreased from 1,742 to 1,447, respectively.

The results of the city center data set are shown in Figure 6. They demonstrate an even larger improvement. At $N_{Hyp} = 500$, the accuracy (MOTA) increased by 12% (71% vs. 83%). Since the environment contains many regions of

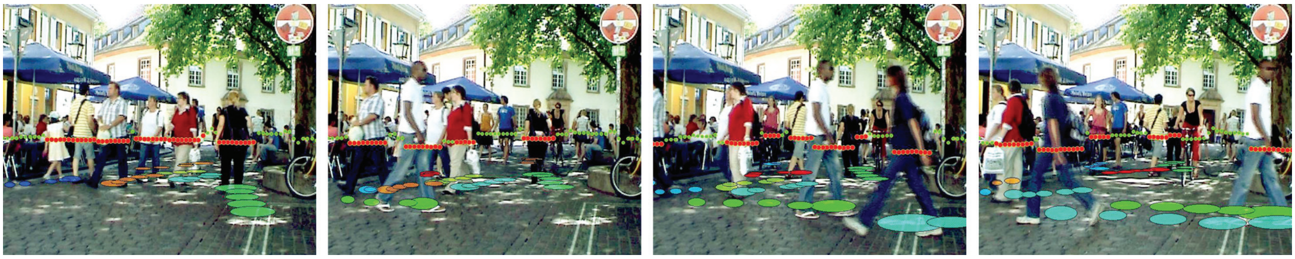

Fig. 7. Four frames of the outdoor experiment carried out in the city center of Freiburg. The images from left to right show the tracking results at time steps $t = 335$, 353, 365 and 381. Laser range measurements are shown as small green dots (background) and small red dots (detected pedestrians). The traces of the observed pedestrians are drawn with colored ellipses.

clutter, the number of false positives decreased substantially (7,106 vs. 3,922). The number of mismatches also dropped from 197 to 125 while the number of misses decreased from 1,992 to 1,552. The 'background learning ability' of the false-alarm layer in the spatial affordance map was particularly appropriate in this data set as the environment contains several person-shaped objects (trees, chairs, trash bins) that led to many false positives from the detector. The fixed-rate approach was not able to cope well with these detection errors and incorrectly created tracks at these locations.

In the diagrams of the three experiments, it can be seen that the number of misses decreased and the number of false positives increased over N_{Hyp}. This behavior is explained by the fact that false alarms are more likely than new tracks. Hypotheses that postulate observations as false alarms receive higher probabilities and can dominate the hypothesis ranking. This can lead to the rejection of lower probability hypotheses at small values for N_{Hyp}, that should have been interpreted as observations of new tracks. With increasing N_{Hyp} more new-track hypotheses survive the pruning process and the number of misses decreases.

The noise in the error plots, such as the number of mismatches, for instance, means that more hypotheses do not always lead to a smaller error, which is counterintuitive. This is due to the pruning strategies in combination with numerical issues in MHT. It follows from the combinatorics of the approach that several hypotheses can have the same probability. If N_{Hyp} happens to prune within such a plateau in the distribution, the outcome of the tracker can become somewhat unpredictable since it depends on the order in which these hypotheses are stored in memory.

In addition to the improvement in tracking performance, the extended tracker is also slightly more efficient. As the new approach makes fewer track-creation errors, it has to maintain fewer tracks on average, especially in regions of clutter. The implementation of our system runs at the sensor frame rate of 12 Hz on a single core of a 2.8 GHz PC with up to 300 hypotheses. With 500 hypotheses, the tracker still runs with 6 Hz.

5.2. Place-dependent motion model

In this section we evaluate the place-dependent motion model from Section 4. A data set was collected in an office environment and divided into a training set and a test set. The training set contains 7,443 frames with 50 person tracks and was used to learn the spatial affordance map (see Figure 8). To learn the walkable-area map, we counted the track confirmation events of the best hypothesis. The test set with 6,971 frames and 28 people tracks was used to compare the model with a constant-velocity motion model under different conditions. The data set was labeled by hand to determine both the ground truth (x, y)-positions of subjects and the true data associations.

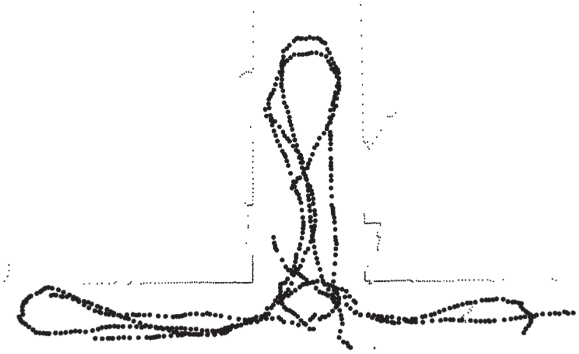

Fig. 8. From experiment 2, 6 (of 50) sample tracks.

To analyze the robustness and accuracy of the new prediction model, we defined, in a first experiment, areas in which target measurements are ignored as if subjects had been occluded by an object or another person. These areas occur at hallway corners and U-turns where people typically maneuver. As the occlusions are only simulated, the ground truth position of the targets are still available. See Figure 3 for sample frames.

For the 28 manually inspected tracks of the test set, the constant-velocity motion model lost a track 12 times while the new model had only a single track loss. Clearly, as a naive countermeasure, one could enlarge the process noise covariance of the constant-velocity motion model to avoid such losses. But in the multi-target case considered here, this leads to enlarged validation gates and increased levels of data association ambiguity. Consequently, the probability distribution over pruned hypothesis trees will be less accurate and lead to a less efficient tracker.

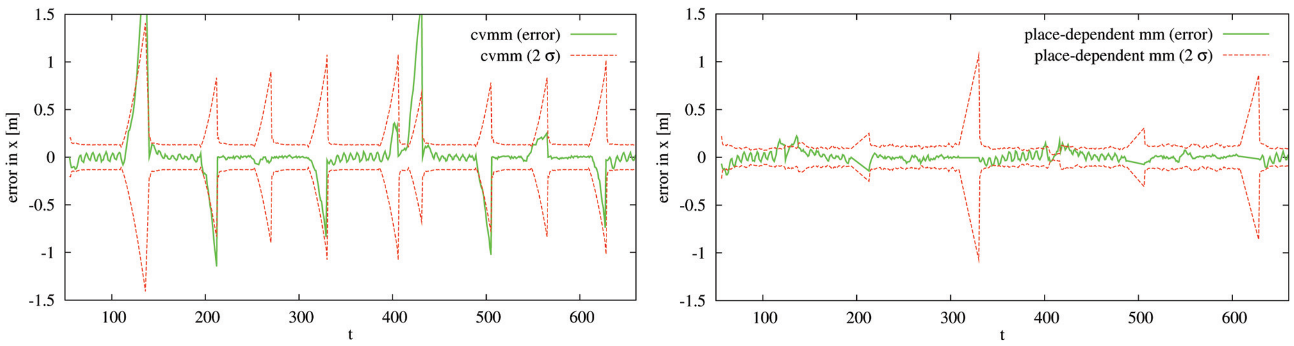

Fig. 9. Estimation error in x of the constant-velocity motion model (cvmm, left) and the place-dependent motion model (right). Peaks correspond to occluded target maneuvers. See also Figure 3, center, which shows the right turn of a person in this experiment. While both approaches are largely consistent from an estimation point of view, the place-dependent model results in an overall smaller estimation error and smaller uncertainties.

As a measure of metric accuracy, the resulting estimation error in x is shown in Figure 9 (the errors in y are similar). The diagram shows smaller estimation errors and 2σ bounds for the place-dependent motion model during most target maneuvers. The predicted covariances do not become boundless during occlusion events (peaks in the error plots) since the shape of the covariance predictions follows the walkable-area map around the very position of the target. Sample situations of this behavior are shown in Figure 3.

In a second experiment, we reduced the observation frequency to 0.5 Hz and we allowed the tracker one second to initialize its targets. The internal cycle time of the tracker was left unchanged at 12 Hz. This setting simulates a very slow data acquisition sensor or the realistic situation of an embedded CPU where people detection runs concurrently with many other processes at a low rate.

The constant-velocity motion model was not able to follow the maneuvering targets and lost all of them as soon as they passed the corner of the hallway. The place-dependent motion model was able to predict the targets around corners as seen in Figure 3 and lost only six of the 28 tracks.

6. Conclusions

In this paper we presented the spatial affordance map for the purpose of extending a people tracker with spatial priors on human behavior. We approached the problem as a parameter estimation problem of a non-homogeneous spatial Poisson process. The model is learned using Bayesian inference from observations of track-creation, confirmation and false-alarm events. It enabled us to overcome the usual fixed Poisson rate assumptions for new tracks and false alarms and to learn a place-dependent model for these events. Finally, we showed that the Poisson process can be seamlessly integrated into the framework of an MHT tracker.

In large-scale experiments in different indoor and outdoor settings, we demonstrated that the extended tracker is significantly more accurate in terms of the CLEAR MOT metric. In particular, the number of track identifier switches was reduced from at least 36% up to several factors. This error is the most relevant metric for a people tracker as it quantifies the ability to keep correct identities over occlusion events and missed detections. The number of false positives dropped by at least 45% while track misses decreased by at least 17%.

The map also allowed us to derive a novel, place-dependent model for predicting the paths of maneuvering targets during lengthy occlusion events. The model is based on a walkable-area map derived from the learned rate function of track-confirmation events and uses an auxiliary particle filter that probes the map at locations of a look-ahead particle. In our experiments, the tracker could follow highly mobile people at an observation frequency as low as 0.5 Hz, clearly outperforming the constant-velocity motion model in terms of track losses.

Notes

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

Funding

This work has been partially supported by the German Research Foundation (DFG; grant number SFB/TR-8).

Conflict of interest

The authors declare they have no conflicts of interest.

References

Arras KO, Grzonka S, Luber M and Burgard W (2008) Efficient people tracking in laser range data using a multi-hypothesis legtracker with adaptive occlusion probabilities. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*.

Arras KO, Martínez Mozos Ó and Burgard W (2007) Using boosted features for the detection of people in 2D range data. In

Proceedings of the International Conference on Robotics and Automation (ICRA). Rome, Italy.

Bar-Shalom Y and Li X-R (1995) *Multitarget-Multisensor Tracking: Principles and Techniques*. Storrs, CT: YBS Publishing

Bernardin K and Stiefelhagen R (2008) Evaluating multiple object tracking performance: the CLEAR MOT metrics. *Journal of Image and Video Processing* 2008: 1–10. DOI: 10.1155/2008/246309.

Best R and Norton J (1997) A new model and efficient tracker for a target with curvilinear motion. *IEEE Transactions on Aerospace and Electronic Systems* 33: 1030–37.

Blackman SS (2004) Multiple hypothesis tracking for multiple target tracking. *IEEE Aerospace and Electronic Systems Magazine* 19: 5–18.

Breitenstein MD, Reichlin F, Leibe B, Koller-Meier E and Gool LV (2009) *Robust tracking-by-detection using a detector confidence particle filter.*

Bruce A and Gordon G (2004) Better motion prediction for people-tracking. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, Barcelona, Spain.

Cox IJ and Hingorani SL (1996) An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18: 138–50.

Cui J, Zha H, Zhao H and Shibasaki R (2005) Tracking multiple people using laser and vision. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Alberta, Canada.

Cui J, Zha H, Zhao H and Shibasaki R (2006) Laser-based interacting people tracking using multi-level observations. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Beijing, China.

Fod A, Howard A and Mataríc M (2002) Laser-based people tracking. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*.

Khan Z, Balch T and Dellaert F (2006) MCMC data association and sparse factorization updating for real time multitarget tracking with merged and multiple measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28: 1960–1972.

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

Kluge B, Köhler C and Prassler E (2001) Fast and robust tracking of multiple moving objects with a laser range finder. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*.

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

Liao L, Fox D, Hightower J, Kautz H and Schulz D (2003) Voronoi tracking: Location estimation using sparse and noisy sensor data. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Mazor E, Averbuch A, Bar-Shalom Y and Dayan J (1998) Interacting multiple model methods in target tracking: a survey. *IEEE Transactions on Aerospace and Electronic Systems* 34: 103–123.

Mucientes M and Burgard W (2006) Multiple hypothesis tracking of clusters of people. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Beijing, China.

Murty K (1968) An algorithm for ranking all the assignments in order of increasing cost. *Operations Research* 16: 682–687.

Pitt MK and Shephard N (1999) Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association* 94: 590–599.

Reid DB (1979) An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control* 24: 843–854.

Rong Li X and Jilkov VP (2003) Survey of maneuvering target tracking. Part I: Dynamic models. *IEEE Transactions on Aerospace and Electronic Systems* 39: 1333–1364.

Schulz D, Burgard W, Fox D and Cremers A (2003) People tracking with a mobile robot using sample-based joint probabilistic data association filters. *International Journal of Robotics Research* 22: 99–116.

Streit R and Luginbuhl T (1995) *Probabilistic Multi-hypothesis Tracking*. Technical Report NUWC-NPT/10/428, Naval Underwater Systems Center, Newport, RI, USA.

Taylor G and Kleeman L (2004) A multiple hypothesis walking person tracker with switched dynamic model. In *Proceedings of the Australasian Conference on Robotics and Automation*, Canberra, Australia.

Topp E and Christensen H (2005) Tracking for following and passing persons. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Alberta, Canada.