Conditional Particle Filters for Simultaneous Mobile Robot Localization and People-Tracking

Michael Montemerlo, Sebastian Thrun, William Whittaker

Abstract—**This paper presents a probabilistic algorithm for simultaneously estimating the pose of a mobile robot and the positions of nearby people in a previously mapped environment. This approach, called the conditional particle filter, tracks a large distribution of people locations conditioned upon a smaller distribution of robot poses over time. This method is robust to sensor noise, occlusion, and uncertainty in robot localization. In fact, conditional particle filters can accurately track people in situations with global uncertainty over robot pose. The number of samples required by this filter scales linearly with the number of people being tracked, making the algorithm feasible to implement in real-time in environments with large numbers of people. Experimental results illustrate the accuracy of tracking and model selection, as well as the performance of an active following behavior based on this algorithm.**

I. Introduction

As robots are deployed in everyday human environments, they will be called upon to perform increasingly interactive tasks. Interaction between humans and robots may occur in a variety of different ways, such as spoken dialog, physical interaction, or the collaborative execution of a task. In order for robots in social environments to be successful, they must be able to both observe and model the behavior of the humans they are working alongside.

Interactive navigational tasks, such as leading, following, intercepting, and avoiding people, require the ability to track human motion. While this paper will concentrate solely on the application of tracking the movement of people in the vicinity of a mobile robot, the results are applicable to a broader class of estimation problems.

The majority of prior approaches to people-tracking have been appearance-based methods. These methods attempt to detect the appearance of people in sensors and track these features over time. Many examples of appearance-based people-tracking have used cameras as the primary sensor [**?**], [**?**], however laser range finders have also been used [**?**]. The accuracy of this approach is limited primarily by the accuracy of feature detection algorithms. In particular, drastic variations in a person's image caused by changes in illumination, viewing angle, and individual appearance make robust detection using vision an extraordinarily difficult problem.

M. Montemerlo and W. Whittaker are with the Robotics Institute at Carnegie Mellon University, Pittsburgh, Pennsylvania. Email: {mmde, red}@ri.cmu.edu

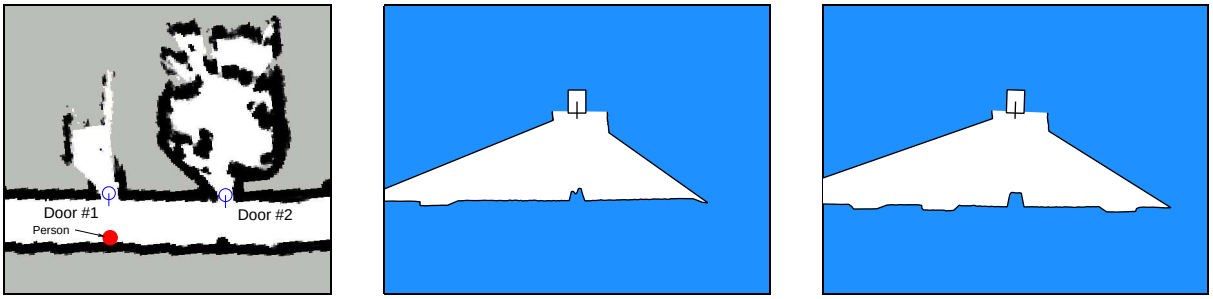
S. Thrun is with the Department of Computer Science at Carnegie Mellon University, Pittsburgh, Pennsylvania. Email: thrun@cs.cmu.edu

Mobile robots operating in fixed environments frequently have maps of their surroundings. By describing the vast majority of objects in the world that are *not* people, maps provide considerable information that can be used to explain individual sensor readings. By comparing the movement of sensor readings that do not correspond with objects in the map against models of human motion, unexpected readings can be used to identify and track people in a very robust manner.

Before maps can be used to categorize the origin of a robot's sensor readings, the pose of the robot relative to that map must be known to some degree. Indeed, localization and map-based people-tracking represent two sides of the same coin. If a robot's exact position in a map is known, determining which sensor readings correspond with objects in the map is a trivial exercise. Conversely, if the sensor readings that correspond with people and other unmapped objects can be filtered out, the true pose of the robot can be determined with maximum accuracy. When the pose of the robot and the positions of people in the map are both unknown, people-tracking and robot localization become a joint estimation problem.

To illustrate this point, consider a mobile robot operating in the map shown in Figure 1(a). When situated in Door #1 facing into the hallway, the robot sees a person and acquires the laser scan shown in Figure 1(b). When the robot is facing out of Door #2, the robot sees a mapped trash-bin and acquires the laser scan shown in Figure 1(c). While the two laser scans look remarkably similar, they represent significantly different hypotheses. A localization algorithm that does not consider the first hypothesis may confuse the person for the trash-bin. A people-tracking algorithm that does not consider the second hypothesis may track the trash-bin as a person.

This paper will present a probabilistic algorithm for simultaneously estimating the pose of a robot and the locations of nearby people in a previously mapped environment. By approximating this joint distribution as a large set of particles representing people locations conditioned upon a smaller set of particles representing robot pose, the expressive power of the joint hypothesis can be exploited in a way that is still computationally tractable. The resulting algorithm, called a conditional particle filter, is robust to sensor noise, occlusion, and uncertainty in localization. Results will demonstrate simultaneous localization and people-tracking in situations with global uncertainty.



(a) (b) (c)

Fig. 1. (a) The robot is in one of two doorways. Is it in front of (b) Door #1 looking at a person? or (c) Door #2 looking at a trash-bin?

II. Mathematical Approach

A. Introduction to Particle Filters

Particle filters have been applied with great success to many real world estimation and tracking problems, as documented by various chapters in [**?**]. Within robotics, common applications involve the localization of mobile robots [**?**], [**?**] or of other mobile entities such as people, from camera images [**?**] and range scans [**?**].

From a mathematical perspective, particle filters estimate the posterior over unobservable state variables from sensor measurements. In the context of the present paper, *state* refers to the pose of the robot (location and orientation) relative to its environment, along with the number and location of people in the robot's proximity. For the sake of the general discussion of particle filters, the total of all those state variables will be denoted by x; further below, this will be made more concrete by including the robot pose and people locations as explicit components of x.

In particular, let the state at time t be denoted by x_t. Particle filters address situations in which this state is not directly observable. Instead, the robot must rely on sensor measurements and information about the controls it executes to infer the posterior distribution over x. Let z_t denote the sensor measurement acquired at time t and u_t denote the control at time t. Thus at time t, two types of information relevant to the state x_t are available to the robot.

$$\begin{aligned} z^t &:= \{z_1, \ldots, z_t\} \\ u^t &:= \{u_1, \ldots, u_t\} \end{aligned} \tag{1}$$

The goal of particle filtering is to estimate the posterior probability over the state variable x at time t:

$$p(x_t | z^t, u^t) \tag{2}$$

This posterior is calculated recursively (see [**?**], [**?**] for a derivation):

$$p(x_t | z^t, u^t) = \tag{3}$$
$$\eta p(z_t | x_t) \int p(x_t | u_t, x_{t-1}) p(x_{t-1} | z^{t-1}, u^{t-1}) \, dx_{t-1}$$

Here η is a constant normalizer. The conditional probability distribution $p(z_t | x_t)$ is a *measurement model* that will be discussed further in Section II-D. Similarly, $p(x_t | u_t, x_{t-1})$ is a *motion model* whose discussion will be postponed to Section II-E. The recursive update equation (3) is equivalent to the well-known Bayes filters, from which Kalman filters and hidden Markov models are easily derived as special cases.

The key idea of the particle filter is to approximate this posterior by set of hypothesized states—called *particles*—which are distributed according to $p(x_t | z^t, u^t)$. Put mathematically, $p(x_t | z^t, u^t)$ is represented by a set of particles

$$X_t := \{x_t^{[i]}\}_{i=1,\ldots,N} \tag{4}$$

where N is the size of the particle set (e.g. $N = 1,000$). It is well-known that such a set of particles X_t can be obtained via the following sampling procedure, which is directly derived from the recursive update equation (3).

$X_t = \emptyset$
for $i = 1$ to N do
  take $x_{t-1}^{[i]}$ from X_{t-1}
  draw $x_t^{[i]} \sim p(x_t | u_t, x_{t-1}^{[i]})$
  calculate (non-normalized) weight $w_t^{[i]} = p(z_t | x_t^{[i]})$
endfor
for $i = 1$ to N do
  draw k with probability $w_t^{[k]} / \sum_{l=1}^{N} w_t^{[l]}$
  add $x_t^{[k]}$ to X_t
endfor

In essence, this procedure utilizes the particle set X_{t-1} to generate a set of particles $\{x_t^{[i]}\}$ that represent the guess at time t after executing control u_t, but before taking into consideration the measurement z_t. Subsequently, the procedure resamples those guesses in proportion to the perceptual probability $p(z_t | x_t^{[i]})$. This simple algorithm has been shown to converge to the desired posterior $p(x_t | z^t, u^t)$ as $N \longrightarrow \infty$.

B. Factored Representations

When particle filters are applied to the problem of people-tracking, the state of the world x is comprised of a set of people locations:

$$x_t = \{y_{t,1}, y_{t,2}, \ldots, y_{t,M}\} \quad (5)$$

Here M denotes the number of people, which is assumed to be known for now. A simple approach to estimating M will be presented in Section II-F.

The general problem with using this state vector lies in its dimensionality. The number of particles required in particle filtering grows exponentially with the size of the state vector, and hence with M. Tracking many people is therefore infeasible with such an approach. The feature-based people-tracking literature usually overcomes this problem by tracking individual people using separate filters [**?**]. The computation in such an approach grows linearly with the number of people M. The underlying mathematical assumption is that the posterior of people locations can be factored as follows:

$$p(y_{t,1}, y_{t,2}, \ldots, y_{t,M}|z^t, u^t) = \prod_{m=1}^{M} p(y_{t,m}|z^t, u^t) \quad (6)$$

Such a decomposition would be mathematically legitimate under two conditions: People move independently, and the robot can reliably identify individual people (i.e., there is no *data association problem*). The first assumption is usually a good approximation. The second is overcome by using a *maximum likelihood* method for the data association; that is, each person observation is assigned to the nearest person track. In this way, conventional feature-based people trackers can reliably track people in a way that scales linearly with M.

C. The Conditional Particle Filter

The total factorization (6) works well for feature-based tracking approaches such as the one in [**?**], but it fails to apply to map-based tracking. Different poses of a robot relative to a map can lead to very different interpretations of sensor measurements, as was illustrated in Figure 1. Thus, the state vector x_t is defined as follows:

$$x_t = \{r_t, y_{t,1}, y_{t,2}, \ldots, y_{t,M}\} \quad (7)$$

Here r_t denotes the robot's pose at time t, and $y_1, \ldots, y_M$ denote the locations of the M people, as above. At fist glance, one might try to factor the posterior just like the feature-based tracking approach:

$$p(x_t|z^t, u^t) = p(r_t|z^t, u^t) \prod_{m=1}^{M} p(y_{t,m}|z^t, u^t) \quad (8)$$

However, this factorization does not work in situations where the robot's pose r_t is uncertain. This is because there are interactions between the robot pose estimate r_t and the people location estimates; depending where the robot is, sensor measurements may be explained by people or by known objects in the map. The fully factored representation (8) will not capture such dependencies and will therefore fail in practice.

The **conditional particle filter** overcomes this problem by using the following representation:

$$p(x_t|z^t, u^t) = p(r_t|z^t, u^t) \prod_{m=1}^{M} p(y_{t,m}|r_t, z^t, u^t) \quad (9)$$

The key difference is that the people pose estimates $p(y_{t,m}|r_t, z^t, u^t)$ are now conditioned on r_t, the robot pose. Thus, our approach factors the desired posterior into a product of a robot pose posterior $p(r_t|z^t, u^t)$ and a product of *conditional posteriors* $p(y_{t,m}|r_t, z^t, u^t)$, which are conditioned on the robot pose r_t. By doing so, dependencies between the robot pose estimate and people location estimates are fully considered, while still maintaining the linear complexity of the basic algorithm in the number of people M.

Conditional particle filters represent both types of posteriors using separate sets of particles. The (unconditional) robot pose posterior $p(r_t|z^t, u^t)$ is represented by a particle set R_t of robot poses $r_t^{[i]}$, just as in plain particle filtering. The conditional distributions $p(y_{t,m}|r_t, z^t, u^t)$ are also represented by particle sets, where each particle set $Y_{m,t}^{[i]}$ is attached to one particular robot particle $r_t^{[i]}$. Put differently, if there are N_r particles representing the posterior of the robot pose, there will be N_r particle sets representing the posterior over a people position conditioned on the robot pose. Each such particle will be denoted $y_{t,m}^{[i,j]}$. The resulting conditional particle filter algorithm is outlined in turn:

$R_t = Y_{1,t} = \ldots = Y_{M,t} = \emptyset$
for $i = 1$ to N_r do
  take $r_{t-1}^{[i]}$ from R_{t-1}
  sample $r_t^{[i]} \sim p(r_t|u_t, r_{t-1}^{[i]})$
  for $j = 1$ to N_y do
    for $m = 1$ to M do
      take $y_{m,t-1}^{[ij]}$ from $Y_{m,t-1}^{[i]}$
      sample $y_{m,t}^{[ij]} \sim p(y_{m,t}|u_t, y_{m,t-1}^{[ij]})$
    endfor
    $w^{[ij]} = p(z_t|r_t^{[i]}, y_{1,t}^{[ij]}, \ldots, y_{M,t}^{[ij]})$
  endfor
  for $j = 1$ to N_y do
    select k with probability $w^{[ik]} / \sum_{l=1}^{N_y} w^{[il]}$
    for $m = 1$ to M add $y_{m,t}^{[ik]}$ to $Y_{m,t}^{[i]}$
  endfor
  $w^{[i]} = \sum_{j=1}^{N_c} w^{[ij]}$
endfor
for $i = 1$ to N_r do
  select k with probability $w^{[i]} / \sum_{l=1}^{N_r} w^{[l]}$
  add $r_t^{[k]}$ to R_t
endfor

Obviously, the resulting number of particles is much larger than in the plain particle filter. However, this approach still can be run in real-time. The primary advantage of this approach is that dependencies between people and robot estimates are fully maintained (a prerequisite for using a map in people-tracking), while different people are tracked using independent filters. The latter property guarantees linear computational complexity of the overall approach.

Four issues must be resolved before map-based people-tracking can be implemented with a conditional particle filter. The form of the *measurement model* and the *motion model*, both described in Section II-A, must be determined. These models describe how the conditional particle filter responds to new sensor observations and actions. A procedure for *data association* must be established, so that new sensor readings can be assigned to the appropriate people filters. Finally, the question of *model selection* must be addressed. For a given robot pose, a procedure for determining the correct number of people M must be determined. These issues will be discussed in the following four sections.

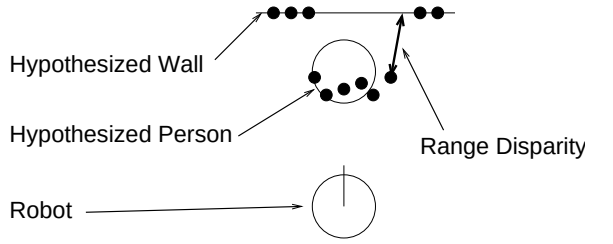


Fig. 2. A slight mismatch between the true position of a person and the hypothesized position leads to a large range disparity. This disparity signifi cantly decreases the probability of the scan.

D. Measurement Model

The form of measurement model $p(z_t|r_t^{[i]}, y_{1,t}^{[ij]}, \ldots, y_{M,t}^{[ij]})$ can have important consequences for the performance of the conditional particle filter for map-based people-tracking. This model characterizes the probability of receiving a sensor reading given a description of the state of the world. In other words, this model compares what the robot actually senses against what it should "expect" to sense given the hypothesized state. Typically, this model is based upon the physics of the real-world sensor being used and its interaction with the environment.

The robot that was used to demonstrate this algorithm used a 2-D laser range finder as its primary sensor. Using the physics-based approach, humans in laser scans can be modeled as approximately cylindrical. The measurement model can be calculated by comparing the actual laser scan with a laser scan ray-traced from the hypothesized robot position in the map. Unfortunately, small differences between the true and hypothesized people positions can cause a large difference in the probability of a given laser scan. Consider the situation shown in Figure 2. A laser measurement, expected to pass by the hypothesized person, actually hits the person. This large disparity in distance causes the range reading, and thus the entire laser scan, to receive very low probability. This lack of smoothness in this measurement model mandates that a larger number of particles be used to accurately represent the posterior distribution over time [**?**].

The sensor model can be made much smoother by calculating probabilities based on disparities in x-y space instead of disparities in range. To calculate the probability of a given robot pose, the laser points are first projected into the world according to the hypothesized pose of the robot. The probability of each point is then computed based on the Euclidean distance between that point and the closest object, be it a person or a occupied map cell. Using this sensor model, the mismatched point in Figure 2 would receive a high probability because it is close to the hypothesized person. The careful construction of a smooth sensor model significantly decreases the number of samples necessary to achieve a particular tracking accuracy.

E. Motion Model

The motion models $p(r_t|u_t, r_{t-1}^{[i]})$ and $p(y_{m,t}|u_t, y_{m,t-1}^{[ij]})$ predict the movement over time of the robot and of people, respectively. The model of robot motion given odometry data is well understood. This model was taken directly from [**?**]. However, no information analogous to odometry is available to describe the actions taken by people in the world. Instead, Brownian motion was used as a model of a person's typical motion. This model predicts that a person can travel in any direction at any time, an assumption that is clearly false. However, using this overly conservative model avoids the need to estimate the velocities and accelerations of people, in addition to their positions. The relatively weak constraint that this model puts on human motion ensures that it will not be violated by people who change direction quickly. This model has shown to work well in practice.

F. Data Association

As a consequence of breaking the estimation of people locations into separate particle filters for each person, each sensor reading must be associated with a particular filter or filters before the weights of each particle can be calculated. If every sensor reading contributed evidence into every people filter, all M filters would track the one most probable person. This association can be a hard assignment, in which each reading is attributed to only one of the person filters or to the map, or it can be a soft assignment, in which each reading can be partially assigned to multiple filters.

When two person filters are far away from each other, the difference between the hard and soft assignment strate-

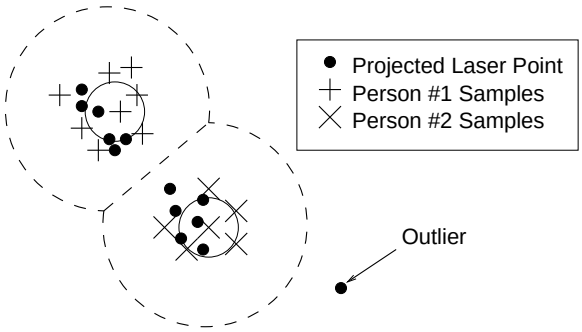


Fig. 3. Laser points are associated with person filters using a modified nearest neighbor rule based on the sample means of each filter. Points more than three standard deviations away from all filters are considered outliers and remain unassigned.

gies is minimal. However, when two filters are very close to each other, experimental results show that soft assignment tends to lump the two filters together. Both filters accept approximately 50 percent assignment of all sensor readings originally associated with the two independent filters. Hard assignment continues to provide good discrimination between the two particle filters even as the means move very close together.

The actual assignment for each laser point is determined using a modified nearest neighbor rule. First, the means and standard deviations of the particles in each person filter are computed. Each laser point is associated with the filter with the closest mean particle, assuming it is within three standard deviations. If the point is more three standard deviations from every filter, it is considered an outlier and not assigned to any class. This procedure is illustrated in Figure 3.

G. Model Selection

The people filters operate on the assumption that the true number of people in the world, M, is known. In practice, choosing an appropriate value of M can be a difficult proposition. People are constantly occluding each other in the robot sensor's field-of-view, especially as they walk very close to the robot. If a person is temporarily occluded by another person, M should not change. However, people also move in and out of the field-of-view of the robot in a permanent way, going around corners and into offices. When a person disappears for an extended period of time, the people filters should respond by decreasing M.

One approach to determining M is to create a prior probability distribution over typical values of M and choose the M at every time step that corresponds with the Minimum Description Length hypothesis. This approach will add a new filter only if it results in a significant gain in the overall probability of the model.

However, this approach requires that multiple instances of every particle filter be run with different values of M. As a result, a great deal of computation is spent on filters that do not represent the true state of the world. This approach can be approximated in a practical manner by examining the associations of laser points and people filters. A cluster of sensor readings that are not associated with any filter indicates that M is too small. A filter that has no laser points associated with it for an extended period of time indicates that M is too large. Experimental results described in the next section illustrate that heuristics based on the laser associations can be used to determine M with high accuracy at a very low computational cost.

III. Experimental Results

A. Tracking and Model Selection Accuracy

The conditional particle filter was tested on a mobile robot equipped with a 2-D laser range finder operating in an office environment. Figure 4(a) shows a typical laser scan given to the algorithm, and Figure 4(b) shows the subsequent state of the conditional particle filter. The people particle filters drawn in the figure correspond to the most likely robot particle. Both people within range of the robot were tracked successfully.

The accuracy of localization and people-tracking were evaluated based on hand-labeled ground truth data captured from a second laser range finder. The standard deviation of the positional error of the robot was approximately 6 cm, and the standard deviation of the positional error of people was less than 5 cm. The mean positional errors of the robot and the people were both less than 3 cm.

The accuracy of model selection was tested on a data set approximately 6 minutes long. Over the course of the run, 31 people passed within the sensor range of the robot. At any given time, up to four people were visible. Of those 31 people, only 3 were not tracked correctly. In one instance, two people entered the robot's field-of-view in close proximity and walked very close to each other. In that situation, both people were tracked incorrectly as a single person.

TABLE I
Conditional Particle Filter Performance

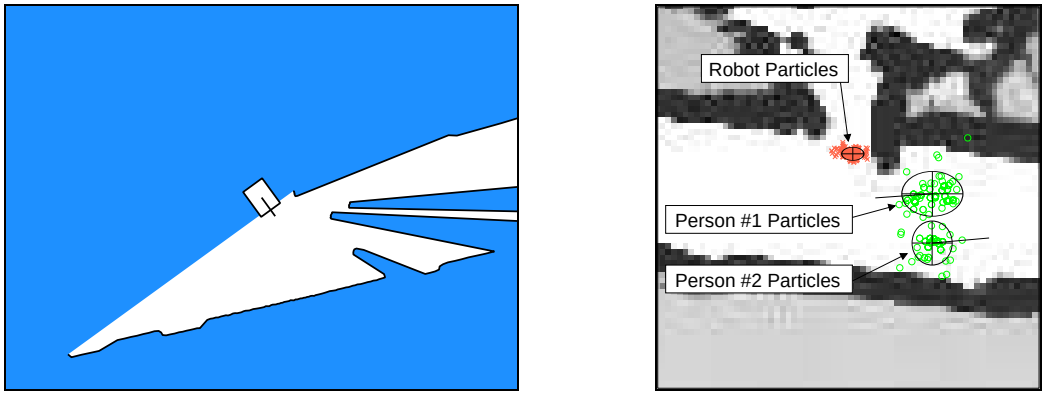
Tracking Accuracy

Robot position - mean error	2.5 cm
Robot position - std. error	5.7 cm
People position - mean error	1.5 cm
People position - std. error	4.2 cm

Model Selection Accuracy

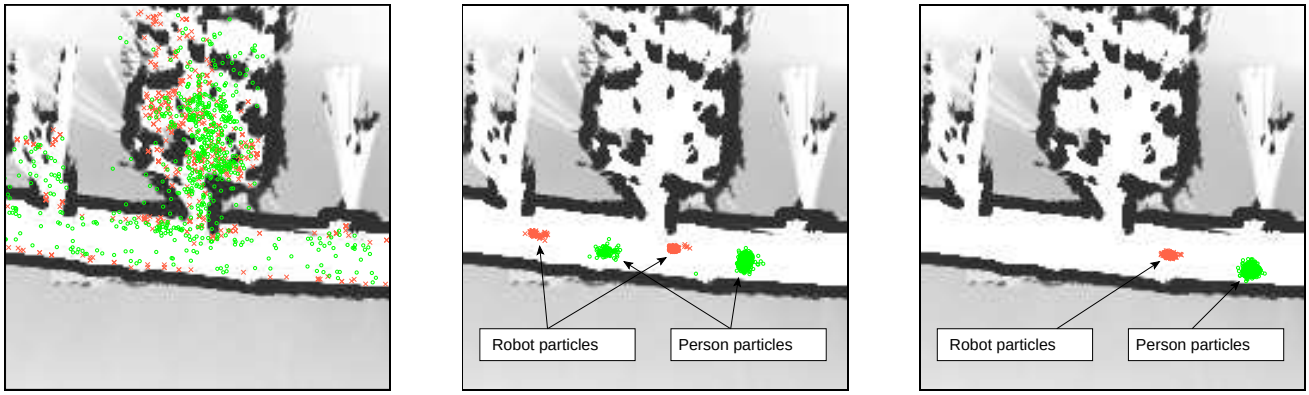
True number of people (cumulative)	31
Model selection errors	3
Model selection accuracy	90%

Model selection was also tested in a more difficult environment, in which the map was not up-to-date. In this run, the robot encountered 11 different people, up to 5 at a



(a) (b)

Fig. 4. (a) Laser scan showing two people near the robot (b) Output of the particle fi lter showing the estimated position of the robot and both people.



(a) (b) (c)

Fig. 5. Evolution of the conditional particle fi lter from global uncertainty to successful localization and tracking

time. All 11 people in this environment were tracked correctly. However, the algorithm also tracked an additional 4 objects. These tracks corresponded with inanimate objects that were not in the map, including a recycling bin in the hallway, a chair, and a closed door that was open when the map was made. In addition to tracking all the people in the incorrect map, the filter was also able to identify inconsistencies between the map and the world. The implications of this will be discussed in Section IV.

B. Global Uncertainty

Figure 4 and Table I illustrate the performance of the conditional particle filter in situations where the position of the robot is relatively well known. The real power of this approach is demonstrated in situations where there is significant uncertainty over robot pose. This commonly occurs during global localization, when a robot is initialized with no information about its position or orientation relative to a map. Figure 5 shows the results of the conditional particle filter during global localization with a single person in the robot's field-of-view. Figure 5(a) shows the state of the particle filter just after initialization, with particles scattered all over the map. After the robot moved a few meters, two modes developed in the distributions of robot and people particles. The two modes, shown in Figure 5(b), correspond to the robot having started from two different doorways in a relatively uniform hallway. Even though there is major uncertainty in the location of the robot, the person is tracked successfully. This is evidenced by the two clusters of people positions moving ahead of the two clusters of robot positions. As the robot moved further down the hallway, sensor evidence eventually disambiguated between the primary hypotheses and converged on the true state of the world, as shown in Figure 5(c).

C. Intelligent Following Behavior

A simple following behavior was implemented using the results of the conditional particle filter. Independent con-

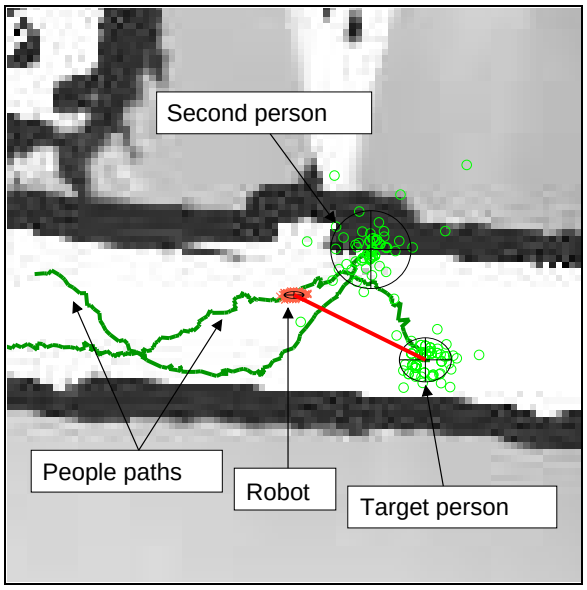


Fig. 6. The following behavior based on the people-tracker output continues to track a person even as that person is occluded repeatedly by a second individual.

trol loops governing the translational and rotational velocity of the robot were based on the relative distance and bearing to the subject being followed. A snapshot of the behavior in operation is shown in Figure 6. The robot was instructed to follow one of two people within range of the robot. A thick line is drawn between the mean robot position and the position of the person being followed. The robot successfully followed this person down the hallway, even as the second person repeatedly walked between the subject and the robot. The robustness of the people-tracker to occlusion enables a very simple control loop to follow a person reliably, even in crowded environments. Movies animating the performance of the following behavior and the conditional particle filter in general are available on the Web at `http://www.cs.cmu.edu/~mmde`.

IV. Discussion

This paper presented the conditional particle filter, an extension to traditional particle filters that breaks high dimensional particles into two sets of lower dimensional particles, one conditionally dependent upon the other. The algorithm was demonstrated in the context of mobile robot localization and people-tracking. The resulting particle filter was able to track people reliably, even in situations with global uncertainty. The algorithm is robust to sensor noise, occlusion, and uncertainty in localization, and the number of samples necessary to implement the filter scales linearly with the number of people in the world.

In addition to tracking people, map-based tracking approaches have the interesting side effect of identify inconsistencies between the map and the world. By observing the positions and velocities of objects over longer periods of time, map-based people-tracking can serve as the basis for a life-long map learning algorithm. Sensor readings that correspond with persistent, stationary, unmapped objects can be used to add these objects into the map. More accurate maps, in turn, will lead to better people-tracking and more accurate navigation.

People-tracking also provides the foundation for numerous high-level robot behaviors. The positions and velocities of people can be used to plan sophisticated actions like intercepting and avoiding people. Robots that use people-tracking information as an input to collision avoidance will be able to actively avoid future collisions, resulting in smoother motion and faster overall navigation.

Acknowledgments

We gratefully acknowledge the Fannie and John Hertz Foundation for their support of Michael Montemerlo's graduate research.