In Proceedings of RoboCup International Symposium, Singapore, June 2010.

Improving People Awareness of Service Robots by Semantic Scene Knowledge

Jörg Stückler and Sven Behnke

University of Bonn
Computer Science Institute VI, Autonomous Intelligent Systems
Roemerstr. 164, 53117 Bonn, Germany
stueckler@ais.uni-bonn.de, behnke@cs.uni-bonn.de

Abstract. Many mobile service robots operate in close interaction with humans. Being constantly aware of the people in the surrounding of the robot thus poses an important challenge to perception and behavior design.

In this paper, we present an approach to people awareness for mobile service robots that utilizes knowledge about the semantics of the environment. The known semantics, e.g., about walkable floor, chairs, and shelves, provides the robot with prior information. We utilize information about the a-priori likelihood that people are present at semantically distinct places. Together with reasonable face heights inferred from scene semantics, this information supports robust detection and awareness of people in the robot's environment. For efficient exploration of the environment for people, we propose a strategy which chooses search locations that maximize the expected detection rate of new persons.

We evaluate our approach with our domestic service robot that competes in the RoboCup@Home league.

1 Introduction

Today's industrial mass production would not be possible without the invention of robots that efficiently and precisely carry out repetitive manufacturing tasks. Just as manufacturing tasks, many of our everyday tasks are monotone, cumbersome or even dangerous. The development of autonomous service robots that one day might relieve humans from these kinds of tasks will thus attain significant importance in the future.

The requirements for service robots differ vastly from those of industrial applications. Manufacturing robots work in an isolated static environment where they fulfill specific tasks. Service robots, on the other hand, need to work in dynamic environments in close interaction with humans. Being constantly aware of the people in the surrounding of the robot thus poses an important challenge to perception and behavior design.

In this paper, we present an approach to people awareness for mobile service robots that utilizes knowledge about the semantics of the environment. The known semantics, e.g., knowledge about the location and properties of furniture like chairs, provides important hints on the presence likelihood and the appearance of persons. In our approach, we exploit such kind of information in various ways. We estimate the presence probability of persons and incorporate semantic knowledge as a prior to the estimation. The presence belief and the knowledge about the appearance of persons enables us to discover false positive detections of persons which is a frequent problem of person detection methods. When the robot needs to explore its environment, we utilize the estimated presence belief to select promising search locations that maximize the expected detection rate of new persons. We evaluate our approach with our domestic service robot that competes in the RoboCup@Home league.

The remainder of this paper is organized as follows: After a brief review of related work in Sec. 2, we outline our method for detecting and tracking multiple persons in the robot's vicinity in Sec. 3. We detail our approach to estimating the presence of persons and how to improve people awareness by semantic scene knowledge in Sec. 4. In Sec. 5 we introduce an efficient method to explore the environment for new people. We evaluate our approach in experiments in Sec. 6.

2 Related Work

Tracking people with laser range finders is a well studied topic in mobile robotics (e.g., [1, 2]). Many approaches detect and track legs of people and fuse this information in a multi-hypothesis tracker [3].

The computer vision community developed a variety of methods for tracking multiple persons with camera systems. For statically mounted cameras (e.g., [4, 5]), background subtraction can be applied to improve the robustness of tracking. When the camera moves (as in [6–9]), subtracting background is no longer possible. Instead, robust person detectors are required that provide stable information for tracking.

Some approaches use spatial and semantic information to improve tracking robustness. In [7], information about tracked objects is fed back as semantic information into visual odometry to improve the robustness of the overall multi-person tracking system. Luber et al. [10] learn Poisson process models of the occurence of persons throughout the environment. They demonstrate improved tracking accuracy with their model compared to standard multi-hypothesis tracking approaches. We propose to use semantic scene knowledge derived from the environment to increase the robustness of people awareness.

3 Continuous People Awareness

For detection and tracking of multiple persons we combine measurements from two laser range finders (LRFs) and a camera. With the LRFs we continuously detect and keep track of possible persons. One LRF is mounted shortly above the ground at a height of 24 cm and detects legs of people. We additionally detect torsos of people with a second LRF at a height of approx. 80 cm.

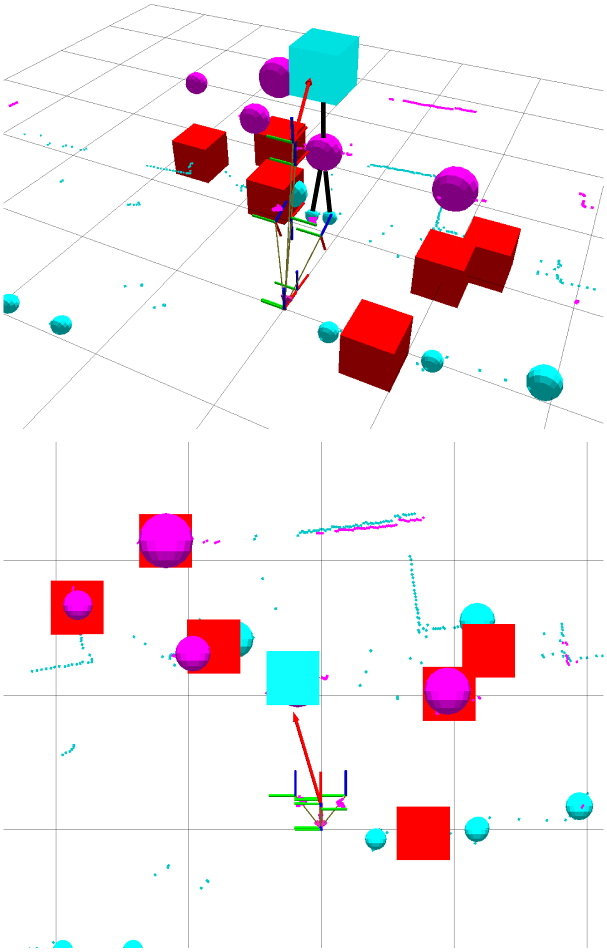


Fig. 1. Persons are detected as legs (cyan spheres) and torsos (magenta spheres) in two laser range scans (cyan and magenta dots). The detections are fused in a multi-hypothesis tracker (red and cyan boxes). Faces are detected with a camera mounted on a pan-tilt unit. We validate tracks as persons (cyan box) when they are closest to the robot and match the line-of-sight towards the face (red arrow). From the projection of the track position onto the face direction we also determine the face height.

We fuse person detections from both LRFs in a multi-hypothesis tracker. We estimate position and velocity of each hypothesis by Kalman filters (KFs). In the KF prediction step, we use odometry information to compensate for the motion of the robot. The tracks are corrected with the observations of legs and torsos.

We use the Hungarian method [11] to associate each torso detection in a scan uniquely with existing hypotheses. In contrast, as both legs of a person may be detected in a scan, multiple leg detections may be assigned to a person hypothesis. Only unassociated torso detections are used to initialize new hypotheses. Spurious tracks with low detection rates are removed.

Of course, the shape of legs and torsos in laser range scans are not discriminative enough, such that parts of the environment cause false positive detections. For this reason, we verify that tracks correspond to persons through a second sensor modality that provides complementary information: We detect frontal and profile views of faces in camera images using the Viola and Jones [12] algorithm. Additionally, we detect upper bodies with a method based on Histograms of Oriented Gradients [13].

The camera is mounted on a pan-tilt unit in a height of approx. 1.5 m. Since the LRFs have a larger field-of-view than the cameras, we implemented an active gaze strategy. We find the correspondence of detected faces and upper bodies with possible person tracks by determining the matching track on the line-of-sight of the detected face that is closest to the robot. We also measure the height of the detected face by projecting the track position onto the face direction. We test that the face has a reasonable height in a later processing stage. Fig. 1 illustrates our approach to continuous people awareness (CPA) with an example.

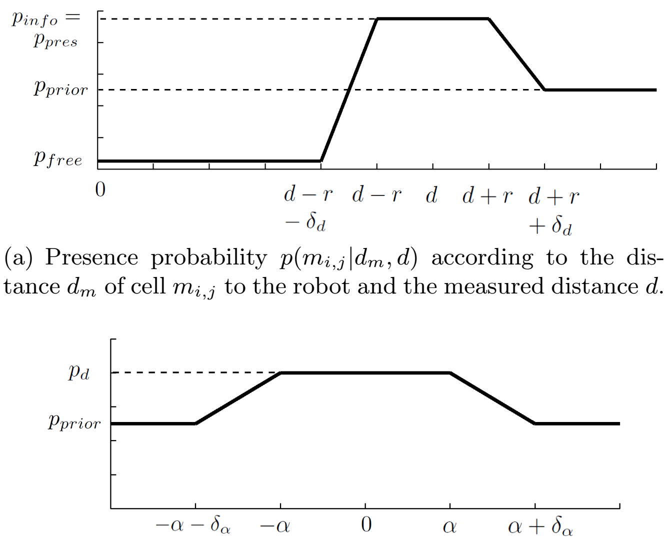
4 Utilizing Semantic Scene Knowledge for Estimating People Presence

Approaches to person tracking are prone to false detections. Low false positive rates in person detection can hardly be achieved, even with more powerful vision-based approaches to person recognition. Thus, we propose to use prior semantic scene knowledge to further increase the robustness of people awareness.

In our approach, we estimate the probability of the presence of people. Similar to the well-known occupancy grid mapping approach to 2D environment mapping, we estimate person presence in a two-dimensional discretized representation of the environment. From semantic annotations of the environment, e.g., walkable floor, chairs, and shelves, we infer prior information on the presence of people. The semantics also provides us with the valid height range of faces. The prior knowledge and the estimated presence likelihood helps us to decide when tracks are falsely detected as persons.

4.1 Estimating People Presence

Our CPA provides the position of persons in the robot's vicinity. Where no person is detected, LRFs measure distance and bearing to the environment structure and thus provide negative information on the presence of people along the measurement. From these measurements and the pose of the robot, we estimate the presence likelihood of people in the environment.



(a) Presence probability $p(m_{i,j}|d_m, d)$ according to the distance d_m of cell $m_{i,j}$ to the robot and the measured distance d.

(b) Presence probability $p(m_{i,j}|\phi_m)$ according to the angular difference ϕ_m of cell $m_{i,j}$ to the measurement.

Fig. 2. Components of the inverse measurement model.

We discretize the environment into a 2D grid. For each cell $(i, j) \in \mathbb{N} \times \mathbb{N}$ in the grid, we estimate the belief of the presence of a person at time t

$$bel_t(m_{i,j}) := p(m_{i,j}|s_{1:t}, z_{1:t})$$

from the poses $s_{1:t}$ of the robot and the positive and negative measurements $z_{1:t}$ up to time t. We assume independence between individual cells, which enables us to estimate the posterior $p(m|s_{1:t}, z_{1:t})$ over the complete map by estimating presence for each cell individually.

Following basically the same derivation as for occupancy grid mapping [14], we arrive at the recursive update scheme in log-odds form:

$$l_t(m_{i,j}) = l_{t-1}(m_{i,j}) + \log\left(\frac{p(m_{i,j}|s_t, z_t)}{1 - p(m_{i,j}|s_t, z_t)}\right) - \log\left(\frac{p(m_{i,j})}{1 - p(m_{i,j})}\right),$$

where we define

$$l_t(m_{i,j}) := \log\left(\frac{p(m_{i,j}|s_{1:t}, z_{1:t})}{1 - p(m_{i,j}|s_{1:t}, z_{1:t})}\right).$$

Here, $p(m_{i,j}|s_t, z_t)$ is the inverse sensor model and $p(m_{i,j})$ is the prior presence probability of cell (i, j).

The inverse measurement model is composed of two parts: The first component $p(m_{i,j}|d_m, d)$ determines the presence probability depending on the distance d_m of cell $m_{i,j}$ to the robot and the measured distance d (cf. Fig. 2(a)).

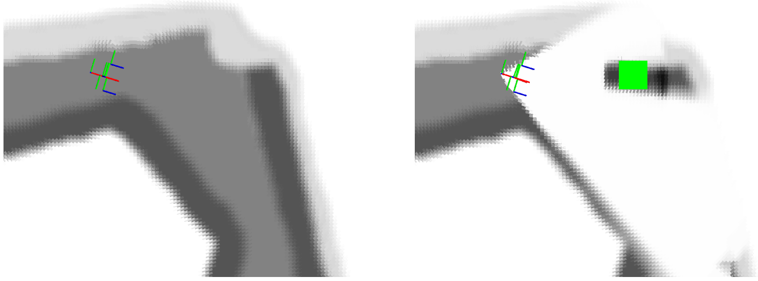


Fig. 3. A person is detected and included into the knowledge base (green square). We update the presence belief (black for probability 1) with positive measurements of the detected person and negative measurements of the environment structure.

We model high likelihood $p_{info} = p_{present}$, when the cell lies within the influence region of a person, which we denote size radius r. For measurements of the environment structure, the presence probability $p_{info} = p_{free}$ is small within an influence radius r. For cells behind the distance $d_m \geq d + r + \delta_d$ we cannot measure, and thus we model prior probability p_{prior}. Cells in front of the robot and with a distance $0 \leq d_m \leq d - r - \delta_d$ are not occupied by persons. Distance δ_d is a small interpolation range to model uncertainty in the measured distance.

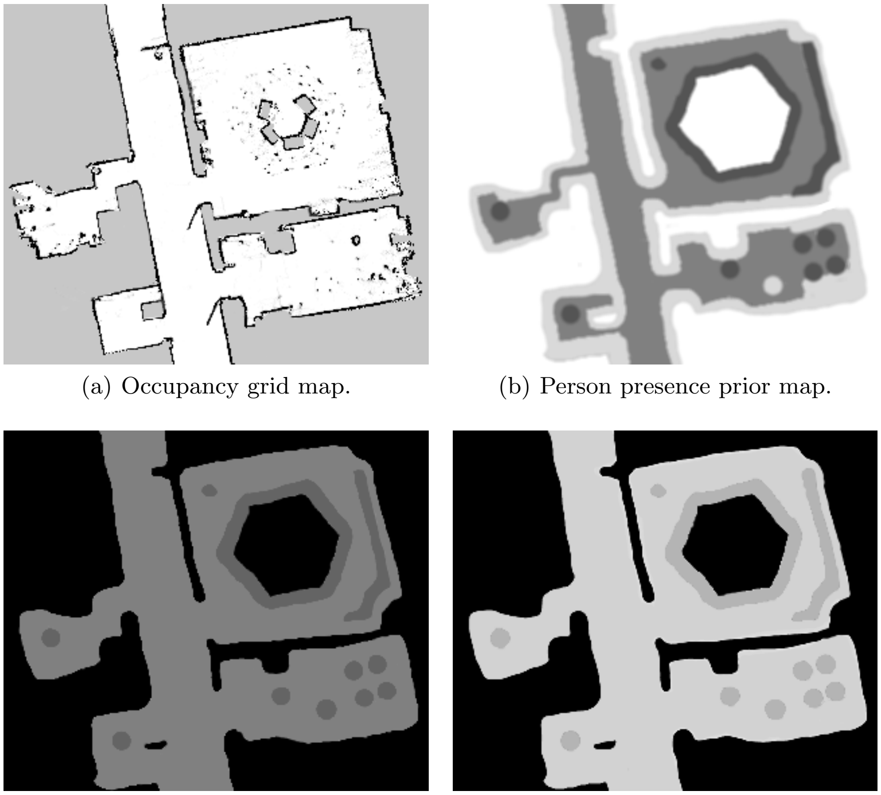
The second part $p(m_{i,j}|\phi_m)$ models influence width and uncertainty along the orthogonal direction to the line of sight. We consider the angular deviation ϕ_m between the cell and the measured position (cf. Fig. 2(b)). In this model, the angular range α is determined from the width of the person or the influence range of a beam. Within this range, the presence likelihood is determined by the first component $p_d := p(m_{i,j}|d_m, d)$. Beyond the angular deviation $|\phi_m| \geq \alpha + \delta_\alpha$ the measurement bears no information for the cell and thus we model prior probability p_{prior}. Again, the interpolation range δ_α models uncertainty in the measurement.

Finally, we consider dynamic changes in the environment by unlearning the acquired information about the presence of people. To implement this exactly, one would require to store a temporal history of measurement updates to each grid cell. As this is not feasible, we approximate unlearning with a small temporal decay towards the prior cell probability.

4.2 Maintaining Person Knowledge

When new persons are found, we represent their size, location, and face height in a knowledge base. Then, the presence map is updated from positive and negative measurements. Fig. 3 illustrates this process in an examplary situation.

While persons are tracked, their information is continuously corrected in the person knowledge base. Simultaneously, the presence map is updated with positive measurements of the tracked persons. We remove persons from the knowledge base, if the presence belief at their location drops below some threshold.



(a) Occupancy grid map.

(b) Person presence prior map.

(c) Minimum face height map.

(d) Maximum face height map.

Fig. 4. Semantic prior maps (b)-(c) for the environment in (a). Darker is lower / more likely.

4.3 Prior Information from Semantic Scene Knowledge

For high-level control of autonomous robots, the semantics of the environment is an important source of information. In the context of people awareness, we propose to exploit the semantics to support the interpretation and disambiguation of sensory data for robust person detection and representation. For example, a chair gives us a hint that persons are more likely to be present at this location than on walkable floor. There, a person typically occupies space for only short durations. In addition, we can use information on the possible postures of people to constrain the detectable heights of faces. Furthermore, we can surely exclude that people reside in walls or furniture like shelves.

We incorporate semantic scene knowledge into our approach in the following way: From semantics, we infer the prior probability of person presence. Additionally, we deduce minimum and maximum face heights throughout the map. We represent this information in 2D grid maps, which enables us to efficiently use the information during presence belief and person knowledge base updates. Fig. 4 shows examples for grid maps of occupancy used for navigation, person presence prior, and minimum and maximum face height.

4.4 Improving People Awareness by Semantic Priors

With semantic priors we can improve the robustness of our people awareness approach. The prior probability on person presence can be directly incorporated into our method for person presence estimation. We use the person presence prior to initialize the posterior estimate. Furthermore, applying Bayes rule to the inverse measurement model, we see that

$$p(m_{i,j}|s_t, z_t) = \eta \; p(z_t|s_t, m_{i,j}) \; p(m_{i,j}).$$

We assume an uninformed prior probability of $p(m_{i,j}) = \frac{1}{2}$ in our previous inverse measurement model and seek to obtain the semantic inverse measurement model by replacing it with the semantic prior $p_{sem}(m_{i,j})$. Unfortunately, the prior is also contained in the normalization factor $\eta = \sum_m p(z_t|s_t, m) \; p(m)$. However, by substituting $p(z_t|s_t, m_{i,j})$ with $\frac{2}{\eta} \; p(m_{i,j}|s_t, z_t)$ in

$$\begin{aligned} p_{sem}(m_{i,j}|s_t, z_t) &= \eta' p(z_t|s_t, m_{i,j}) \; p_{sem}(m_{i,j}), \text{ and} \\ p_{sem}(\neg m_{i,j}|s_t, z_t) &= \eta' p(z_t|s_t, \neg m_{i,j}) \; p_{sem}(\neg m_{i,j}), \end{aligned}$$

and exploiting the fact that $p_{sem}(m_{i,j}|s_t, z_t)$ is a probability measure over a binary variable, we find

$$p_{sem}(m_{i,j}|s_t, z_t) = \frac{p(m_{i,j}|s_t, z_t) \; p_{sem}(m_{i,j})}{p(m_{i,j}|s_t, z_t) \; p_{sem}(m_{i,j}) + (1 - p(m_{i,j}|s_t, z_t)) \; (1 - p_{sem}(m_{i,j}))}.$$

The prior strongly biases belief estimation. At locations with very low prior probability, spurious person detections increase the person presence belief insignificantly. On the other hand, detections in a-priori highly probable areas will faster converge to high presence probability.

We further exploit semantic priors to reject falsely detected persons in regions with low presence probability. Low prior probability in such regions supports the robustness against false detections, as the posterior belief is hardly increased from spurious person detections. Finally, we validate the face height of detections using semantic priors.

5 Exploiting Semantics for Efficient People Search

So far, our approach robustly tracks multiple persons in the robot's surrounding. It estimates a belief of person presence in the viewed environment and also maintains knowledge of found people. Typically, a mobile service robot operates

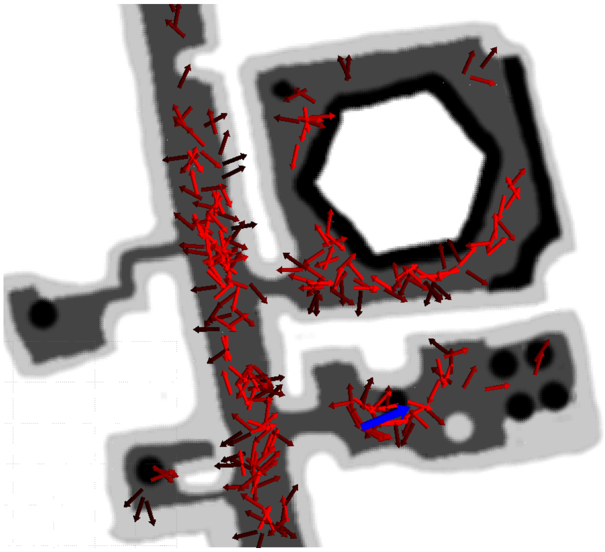


Fig. 5. For efficient exploration, we randomly draw $N = 400$ poses from a normal distribution around the current pose. Among these poses (small red arrows) we select a new exploration pose (big blue arrow) that maximizes the expected detection rate (increasing from dark to bright).

in an environment which it cannot survey from a single location. When the robot further requires awareness of people in the complete environment, an efficient exploration strategy needs to be devised.

For this purpose, we propose an exploration strategy that utilizes semantic prior knowledge. Based on the presence belief of people, the robot selects promising exploration views in which the rate of detecting new persons is maximized.

We sample N random poses from a normal distribution around the current pose of the robot. From these poses we select the one that achieves best expected detection rate in the viewed area. We define the expected detection rate $\mathbf{E}_{\mathcal{M}}(D)$ in a map region $\mathcal{M}$ as

$$\mathbf{E}_{\mathcal{M}}(D) = \sum_{m_{i,j} \in \mathcal{M}} \left(p(D|m_{i,j})\ bel(m_{i,j}) + p(D|\neg m_{i,j})\ bel(\neg m_{i,j}) \right),$$

where $bel(m_{i,j})$ is the current belief estimate for the presence of a person in cell (i,j) and $p(D|m_{i,j})$ is the detection probability given the presence of a person in a cell.

We determine the viewed area $\mathcal{M}$ by ray-casting in the environment map. To prevent search at places where persons have already been found, we exclude regions that are occupied by known persons in the knowledge base. In this way, we measure the expected detection rate of new persons. Combined with our semantic prior, the robot prefers to explore regions with high expected detection rate of new persons.

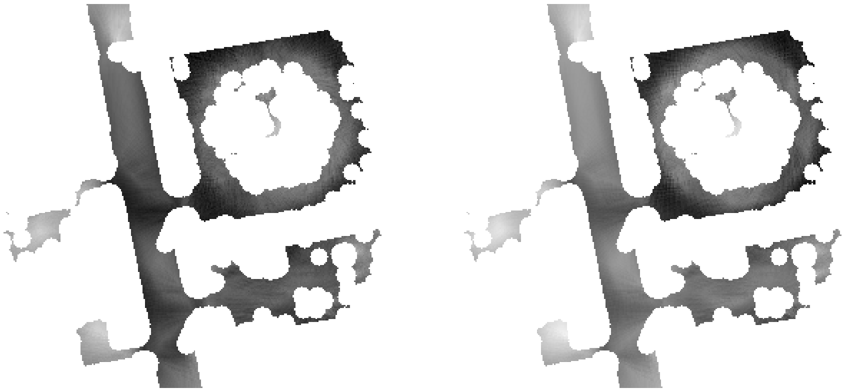


Fig. 6. Expected detection rate in discretely samples poses without (left) and with (right) semantic prior. For each position the best orientation is shown. With semantic prior the expected detection rate is high for poses that view regions with predominantly high presence likelihood. Without prior, the number of cells under the pose's view solely determines the detection rate measure.

6 Experiments

We first present exemplary results for our exploration strategy. Fig. 6 visualizes the effect of the semantic prior on the expected detection rate. We sample poses at equidistant positions with a spacing of 5 cm and in $\frac{\pi}{8}$ orientation steps. At each location we determine the maximum value over orientations and normalize all values to the interval $[0, 1]$. Without prior, the number of cells under the pose's view solely determines the detection rate measure. When we add the semantic prior, the expected detection rate is high for poses that view regions with predominantly high presence likelihood. By utilizing semantic information, our exploration approach concentrates on those regions with high detection rate of new persons.

We also evaluate our exploration approach with our domestic service robot that competes in the RoboCup@Home league. We conduct an experiment in our test environment (cf. Fig. 7). Three persons are placed at the locations depicted in the upper left of Fig. 7. Person presence prior and face height maps have been manually designed to represent the semantics of the environment. A timeline of the events in the experiment is given on the bottom of Fig. 7. The robot drives to 16 exploration views. It finds all three persons in ca. 909 seconds. Three false positive detections are rejected: the first one due to an invalid face height of approx. 0.55 m, the others as they are located at places with low presence estimate.

We successfully applied continuous people awareness during the RoboCup GermanOpen 2010 in Magdeburg. In the Who-Is-Who test, the robot detected and identified all 5 persons that either sat or stood in an apartment-like environment.

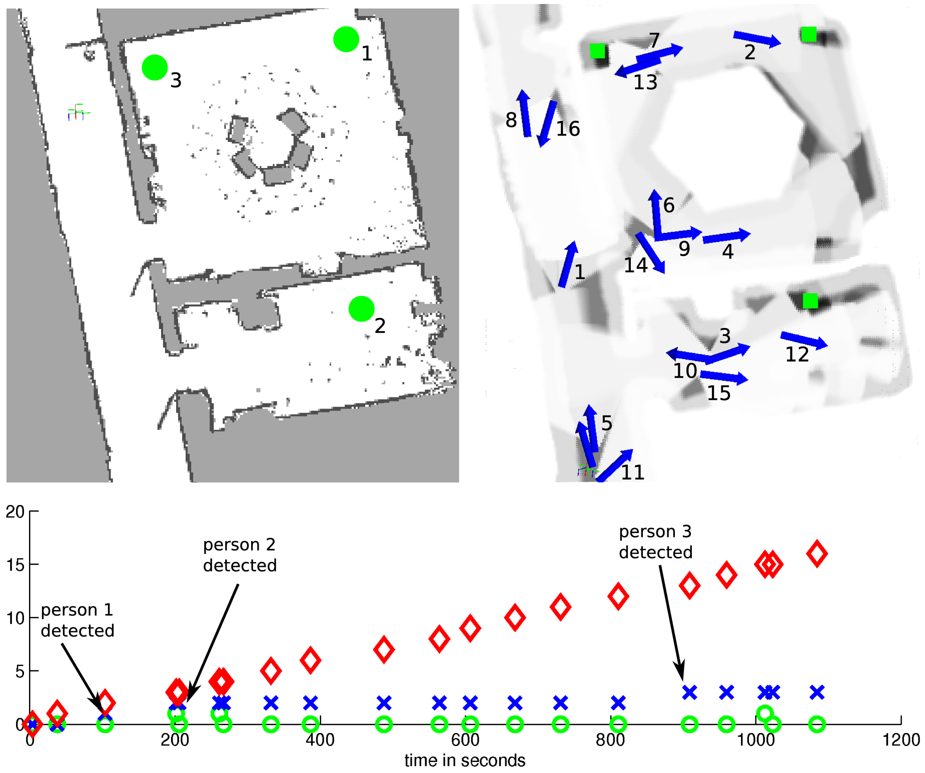


Fig. 7. Top left: Environment of the exploration experiment with three persons (green circles). Top right: Final person presence belief and locations in the knowledge base (green boxes). The blue arrows indicate exploration poses. Bottom: Event timeline of the experiment showing visit time of exploration poses (red diamonds), number of detected persons (blue crosses), and number of represented false detections in the person knowledge base (green circles).

7 Conclusions

We propose an approach to people awareness for mobile service robots that exploits knowledge about the semantics of the environment. From known semantics, we model the prior probability of the presence of persons and represent it in a 2D grid. We estimate the presence probability of persons recursively and incorporate the semantic prior into the estimation. Furthermore, we extract valid face height ranges from semantics.

During person search, we reject false detections in unlikely regions and with invalid face heights. We also propose a search strategy for people that maximizes the expected detection rate in the robot's view. We demonstrate our approach with our domestic service robot that competes in the RoboCup@Home league.

Currently, our approach is limited by the person detection methods we employ. The use of more generic person detection schemes could further improve robustness and detection rate.

In the current system the user provides semantic annotations and interpretations in the form of priors and face height ranges. It is an interesting line of research to recognize objects and places, and to learn semantic prior models by observation of people at such spatial entities.

References

1. Dirk Schulz, Wolfram Burgard, Dieter Fox, and Armin B. Cremers. People tracking with a mobile robot using sample-based joint probabilistic data association filters. *International Journal of Robotics Research*, 22, 2003.
2. Kai Arras, Slawomir Grzonka, Matthias Luber, and Wolfram Burgard. Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Pasadena, CA, USA, 2008.
3. Donald B. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24(6), 1979.
4. Jerome Berclaz, Francois Fleuret, and Pascal Fua. Robust People Tracking with Global Trajectory Optimization. In *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 744–750, 2006.
5. Oswald Lanz. Approximate bayesian multibody tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(9):1436–1449, 2006.
6. Dariu M. Gavrila and Stefan Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *Int. Journal of Computer Vision*, 73(1):41–59, 2007.
7. A. Ess, B. Leibe, K. Schindler, , and L. van Gool. A mobile vision system for robust multi-person tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.
8. Kenji Okuma, Ali Taleghani, Nando de Freitas, James J. Little, and David G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *Springer Lecture Notes in Computer Science (LNCS)*, 2004.
9. Bo Wu and Ram Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266, 2007.
10. Matthias Luber, Gian Diego Tipaldi, and Kai O. Arras. Spatially grounded multi-hypothesis tracking of people. In *Proceedings of the ICRA Workshop on People Detection and Tracking*, 2009.
11. Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1):83–97, 1955.
12. Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001.
13. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886–893, June 2005.
14. Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. MIT Press, 2005.