Bayesian Fusion of Ceiling Mounted Camera and Laser Range Finder on a Mobile Robot for People Detection and Localization

Ninghang Hu[1], Gwenn Englebienne[1], Ben Kröse[1,2]

[1] University of Amsterdam
[2] Amsterdam University of Applied Science

Abstract. Robust people detection and localization is a prerequisite for many applications where service robots interact with humans. Future robots will not be stand-alone any more but will operate in smart environments that are equipped with sensor systems for context awareness and activity recognition. This paper describes a probabilistic framework for the fusion of data from a laser range finder on a mobile robot and an overhead camera fixed in a domestic environment. The contribution of the framework is that it enables seamless integration with other sensors. For tracking multiple people it is possible to use a probabilistic particle filter tracker. We show that the fusion improves the results of the individual subsystems.

1 Introduction

As the baby boom generation is coming to retirement age, the number of elderly citizens over 60 years of age is expected to grow further to a proportion of 1 out of 3 by the year 2030. Alongside this growth in the elderly population, we face short and long-term labor shortages, especially in the health-care sector. Robots may offer a solution for making elderly care affordable by using them for physical [9], cognitive [12] or social [16] support. All these studies share a common foundation that the robots interact intensively with humans, and locations, of both the person and the robot, are estimated robustly.

Sensing systems for robot localization or people localization are usually mounted either on the robot or are fixed in the environment. In this paper we describe a probabilistic framework for the *fusion* of data from robot and fixed sensors. Here we restrict ourselves to a laser scanner on the robot and an overhead camera fixed in the room. The contribution of our work is that by mapping all information into a probabilistic model, the system can be easily extended with other sensors such as multiple cameras or RGB-D cameras, and is robust to the absence of sensors.

In the next sections we will describe related work and systems. In section 4 our own model will be introduced. The sections after that will describe the likelihood functions for the laser on the robot and the camera in the room. In section 7 we describe the experiments and results. We conclude with a discussion on the method and results.

2 Related Work

There is a long tradition of research in the field of people detection and localization in robot applications. Many studies concentrate on people detection using the sensors on the mobile robot. Relatively simple sensors such as laser range finders were used for detection and localization [11, 15]. People are extracted from range data as single blobs or found by merging nearby point clusters that correspond to legs. Probabilistic techniques such as multi-hypothesis trackers are used for tracking multiple objects [1].

Instead of using the laser range systems on the robots, vision systems have also been used for people detection. Since robot-mounted cameras are moving, the detection cannot be based on background modeling methods, and local characteristics such as color histograms or local features have been used [14, 17]. To make detection more robust, the fusion of different modalities of robot sensors is suggested. Leg detection by laser range finders in combination with face detection has shown to be more robust than individual modalities [10, 2]. In [18], Viola-Jones type of visual detectors are used to recognize body parts and are combined with laser range data.

However, future robots will operate in smart homes that are equipped with sensors, and it seems obvious to use these sensors also for person detection. One advantage is that the system may be more robust: noise or deviations in a sensor may be detected and corrected. Another advantage is that the robot does not need to keep monitoring the persons all the time. The robot may be required to finish other tasks from time to time, rather than allocating its resources to the task of tracking each person all the time.

Person tracking systems that are mounted in domestic environments are usually based on vision systems, although there are some exceptions using laser range finders [8] or speech [6]. Overhead cameras are often used, which are usually mounted very high, and have a very wide angle of view, covering most of the areas in the room. Since they look down from above, it turns out that human users are less likely to be occluded compared with cameras mounted on the side. An application in a kids playroom is given in [3].

In our set-up we combine an overhead camera with the laser range finder on the robot. In order to have a sound probabilistic framework we build on the approach of [7], who uses a probabilistic foreground segmentation with a template based detection. The result is a posterior distribution on the locations of the persons in the room. This is combined with a distribution based on the laser range finder.

3 System Overview

Our proposed system is used to detect and localize the elderly people in chores of robot home-care. With our system, the robot is able to obtain accurate locations of the users in the room, and thereby it can interact with the human users precisely. The robot we use possesses multiple on-board sensors, including a Kinect camera, a stereo camera, and a laser range finder.

In the recent work, most of the robots are designed for following the targets. These approaches, therefore, require that the users are always in the range of the robot sensors. In the case of home care, however, the robot moves around in the room to execute a variety of tasks, and at some points the robot sensors will loss the track of the human user, e.g. the robot is asked to get an object that is in an opposite direction to the user. To overcome such a problem and enable continuous human localization, we adopt an ambient camera and mount it on the ceiling of the room. The advantage of the ambient camera is twofold: (1) that it gives a top view of the whole room, and (2) that people in the room are less likely to be occluded compared with the robot cameras. Since it covers the whole area of the room, the ceiling-mounted camera is able to localize persons continuously when the users are present in the room, so that when the robot sensors fail to detect the users, the ambient camera is still able to report the correct location to the system. Besides, the robot sensors and the ambient camera observe the persons from different directions, giving complementary cues for the human detection and localization. The fusion system can, therefore, obtain a better estimate of the location of the users compared with the approaches using single modality.

To combine the robot sensors and the ambient camera, we propose a Bayesian fusion framework. Next, we formulate the problem and introduce our fusion framework.

4 Probabilistic Fusion Framework

The Bayesian approach provides an elegant way of fusing between different sensor sources as well as dealing with noise and uncertainty in sensor measurements [13].

Assume I_R is the observed data from the robot sensor, and I_C is observed from the ambient camera, *i.e.* the overhead camera. Given I_R and I_C, we aim to find a robust estimation of the location of multiple persons L_H, the location of the robot L_R, and the orientation of the robot θ_R. In the context of a Bayesian framework, the posterior distribution $P(L_R, L_H, \theta_R | I_R, I_C)$ is the target we would like to know by the end.

Using the Bayesian Theorem, the posterior probability can be derived as

$$P(L_R, L_H, \theta_R | I_R, I_C) \propto P(I_R, I_C | L_R, L_H, \theta_R) P(L_R, L_H, \theta_R) \tag{1}$$

where $P(L_R, L_H, \theta_R) = p(L_R)p(L_H)p(\theta_R)$ is the prior distribution that is known before the sensory data is observed. These priors can be estimated either from separate training data, or from prior knowledge of the problem. In our case, we simply assume a uniform distribution over the ground area of the floor, and a uniform distribution over the angles of the orientation. $P(I_R, I_C | L_R, L_H, \theta_R)$ is the likelihood.

By assuming I_R and I_C are measured independent with separate sensors, and I_C is not dependent on the rotation of the robot θ_R, the likelihood probability of Equ. 1 can be decomposed as

$$P(I_R, I_C | L_R, L_H, \theta_R) = P(I_R | L_R, L_H, \theta_R) P(I_C | L_R, L_H) \tag{2}$$

where $P(I_R | L_R, L_H, \theta_R)$ is the likelihood of generating the image I_R given the combination of L_R, L_H, and θ_R, while $P(I_C | L_R, L_H)$ represents the likelihood of the ambient camera that generates the observation I_C.

Again, our goal is to find the optimal combination of L_R^*, L_H^* and θ_R^* that maximizes the posterior distribution $P(L_R, L_H, \theta_R | I_R, I_C)$, which is a typical MAP problem that can be solved by particle filtering [5].

The camera likelihood $P(I_C|L_R, L_H)$ is used as the proposal distribution to sample particles, and the particles are weighted by the corresponding likelihood of the laser data $P(I_R|L_R, L_H, \theta_R)$. The optimal combination L_R^*, L_H^* and θ_R^* is considered as the particle that holds the highest weight. In a Bayesian framework, however, we find the expectation of the parameter values rather than the most probable value. Therefore, rather than choosing one particle that maximizes the joint distribution, we compute the solution as a weighted sum of all the particles.

The remaining is to compute the two likelihood terms in Equ. 2. In the following two sections, we introduce the methods of estimating the two likelihood items separately. Here, we will focus on modeling the likelihood of the robot sensor. For the camera, we adopt the algorithm from [7].

5 Measuring Likelihood of Robot Sensors

In our data fusion framework, the state is to be estimated is a triplet of $\{L_R, L_H, \theta_R\}$. The likelihood of the robot sensors measures the probability of generating the observation I_C rather than all the observations that can be possibly generated from such a triplet, given such a state triplet, *i.e.* $P(I_R|L_R, L_H, \theta_R)$.

In this paper, we adopt the Laser Range Finder as our robot sensor. The Laser Range Finder scans in a plain and detects the distance to the objects in range. In the context of human localization in a home setting, the detected objects can either be objects that exist in the room or be part of the human in the room. In this paper, we use the background model and the human model, respectively, to model the probability that a region is occupied by either of these two objects. Then we can compute the occupancy map of the room, *i.e.* the probability that the area is occupied by either the background object or by a human.

The occupancy maps are used to estimate the probability of the robot sensor generating a certain set of observations, *i.e.* the robot sensor's likelihood.

5.1 Probabilistic Background Model

To find out what the room looks like in terms of background obstacles, the robot is first driven around to build a background model of the room.

For each time stamp, the robot sensor fires a set of laser beams $l = \{l_1, l_2, l_3, ...\}$. Whenever there is an object in the way, the laser is reflected back to the base and thereby the distance to the background objects is detected. Given the coarse location of the robot, we are able to find the approximate locations of these laser detections. But due to the uncertainty in the location of the robot as well as the noise in laser data, these locations are not fully reliable. Therefore, simply giving a boolean answer to the occupation of the local region is not an elegant solution, and a probabilistic way of modeling the background is required.

In our approach, the ground plane is first discretized into small cells of equal size. We denote k as the index of the cell on the ground plane. Then for each cell k, we aim to estimate the probability that the cell is occupied by a part of the background. Collectively, these probabilities form the background model $P_b(k)$.

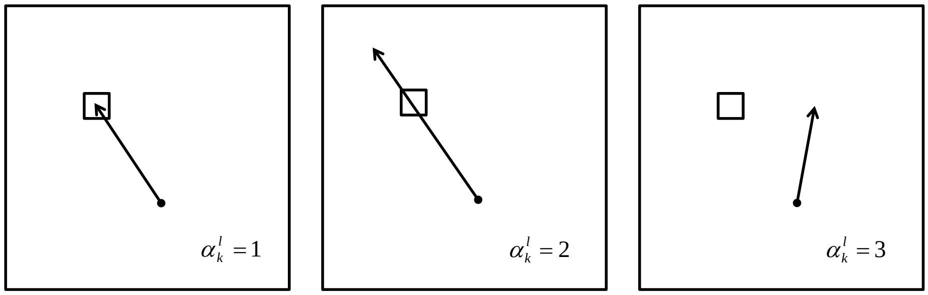

Fig. 1. The relation between a laser beam and a cell can be summarized into three patterns. In the left pattern $\alpha_k^l = 1$, the laser is blocked by the cell, referring that the cell is occupied by certain background objects. The middle pattern $\alpha_k^l = 2$ shows the laser has passed through the cell, indicating the cell is empty. As for the third pattern, however, the laser beam is blocked before it reaches the cell. Therefore, no clue about whether the cell is occupied can be deduced from the third pattern.

In this paper, the background model $P_b(k)$ is measured as the number of times the laser scanner observes an occupied cell normalized by the number of times that the cell is in the range of the laser scanner. To formalize the problem, we define three patterns that can be observed given a scan l and a cell k, see Fig. 1. We use a random variable α_k^l to denote the index of the three patterns. The first pattern refers that the cell k is detected by l as an occupied cell. The second pattern denotes that the cell is observed as an empty cell. As for the third pattern, no information about the cell can be inferred since the cell is either occluded by other background objects that is in front of the cell, or the laser is not fired in the direction of the cell. Therefore, the third pattern does not contribute to the background model while only the first two do. Next, we estimate the background model by

$$P_b(k) = \frac{\sum_l \delta(\alpha_k^l - 1)}{\sum_l \delta(\alpha_k^l - 1) + \delta(\alpha_k^l - 2)} \tag{3}$$

where δ is a Kronecker delta function, and the equation sums over all the lasers that pass through the cell k.

5.2 Learning Human Model

The human model P_h reflects how the human looks like from the robot sensors in the world frame. It is learned by accumulating the laser points that locate in a small region around the center of the person. Each pixel in such a region holds a value indicating the probability that the cell is occupied by the person, *i.e.* a higher value means the cell is more likely to be detected by the robot sensor due to the occurrence of the human.

Similar to training the background model, we learn the human model P_h by calculating the number of laser beams that either have a positive detection at the cell or pass through the cell. Again, we adopt the Equ. 3 for computing the human model P_h.

Given the person locating in cell k, the local human model P_h can be translated into the world frame to generate a human model map $P_h(k)$.

5.3 Occupancy Map

Knowing the background model and the human model, we are able to compute the probability of occupancy for each of the cells on the ground plane. Note that the cell cannot be occupied by both the human and the background obstacle at the same time, therefore the occupancy map is computed as

$$\omega_k = \frac{P_b(k)\tilde{P}_h(k) + P_h(k)\tilde{P}_b(k)}{1 - P_b(k)P_h(k)} \tag{4}$$

where

$$\tilde{P}(k) = 1 - P(k) \tag{5}$$

5.4 Likelihood of Laser Range Finder

The likelihood of the Laser Range Finder denotes the probability of generating the current observation given the state $\{L_R, L_H, \theta_R\}$. I_R represents a vector of the laser range data. Assume I_R contains N independent measurements $\{i_R^1, i_R^2, ..., i_R^n, ..., i_R^N\}$. Suppose the direction of the range measurement i_R^n is defined by θ_R^n. Therefore

$$P(I_R|L_R, L_H, \theta_R) = \prod_{n=1}^{N} P(i_R^n|L_R, L_H, \theta_R^n) \tag{6}$$

L_R and θ_R^n define a robot at the location L_R, and the robot fires a laser beam in the direction of θ_R^n. L_H refers to the location of multiple persons.

Suppose the laser beam i_R^n passes through a set of cells in a straight line, e.g. $\{c_1, c_2, ..., c_{m-1}\}$, and then it detects a certain object at the cell c_m. c_M denotes the maximal range that the laser can reach. See Fig. 2. Then the probability of obtaining a detection at cell c_m rather than the other locations can be computed by

$$P(i_R^n|L_R, L_H, \theta_R^n) = \frac{\omega_{c_m} \prod_{i=1}^{m-1} \tilde{\omega}_{c_i}}{\sum_{j=1}^{M} \omega_{c_j} \prod_{i=1}^{j-1} \tilde{\omega}_{c_i}} \tag{7}$$

Since multiplications of the probabilities can result in very small numbers which lead to floating point overflows, we compute the log-likelihood instead

$$\mathscr{L}(i_R^n|L_R, L_H, \theta_R^n) = \lambda_m - \sum_{j=1}^{M} \lambda_j \tag{8}$$

where

$$\lambda_m = log(\omega_{c_m}) + \sum_{i=1}^{m-1} log(\tilde{\omega}_{c_i}) \tag{9}$$

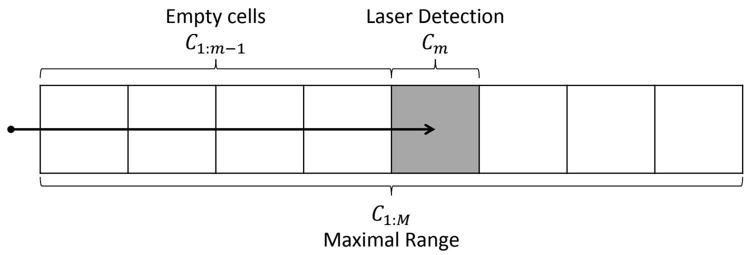

Fig. 2. The laser beam (Arrow) passes through $m-1$ empty cells and finally reaches the cell at C_m. The maximal range of the laser covers M cells.

6 Likelihood of Ceiling Mounted Camera

The likelihood of over head camera is computed the same way as in [7]. Assuming the pixels are independent from each other given the image taken by the ceiling mounted camera, the likelihood $P(I_C|L_R,L_H)$ can be derived as

$$P(I_C|L_R,L_H) = \prod_{n=1:N} P(i_C^n|L_R,L_H) \tag{10}$$

We build a specific polyhedron to model the 3D shape of both the human and the robot. Given the location of the human L_H and the robot L_R, the polyhedrons are projected into the image space, generating a foreground mask $\mathscr{M}$. For each pixel location $P(i_C^n)$ on the image, we look up in the mask and use $\mathscr{M}_n$ to determine whether the pixel is a part of the foreground or background. Then the likelihood can be computed as

$$P(i_C^n|L_R,L_H) = P_f(i_C^n)\mathscr{M}_n + P_b(i_C^n)(1-\mathscr{M}_n) \tag{11}$$

where $P_b(i_C^n)$ is the background model which is learned beforehand using the background images. $P_f(i_C^n)$ is the foreground model, and in our case we apply a uniform distribution over the colors.

7 Experiment and Results

The proposed data fusion framework was evaluated on data collected with a Nomad robot platform and an overhead camera, see Fig. 3. The overhead camera is mounted centrally on the ceiling and gives a panoramic view of the room. The frames that are captured with the camera are highly distorted due to the fish-eye effect. The camera's lens parameters are calibrated with the OpenCV module [4].

On the Nomad robot platform, a Laser Range Finder, a Kinect camera and a stereo camera are mounted on the robot. For the present experiments, we restrict ourselves to test the framework by using the Laser Range Finder, mounted at a height of 20 cm. The robot is remote-controlled and manually driven around in the room. The robot records its odometry information by measuring the rotations of its two wheels. The odometry data are then adopted for generating the orientation and location of the robot.

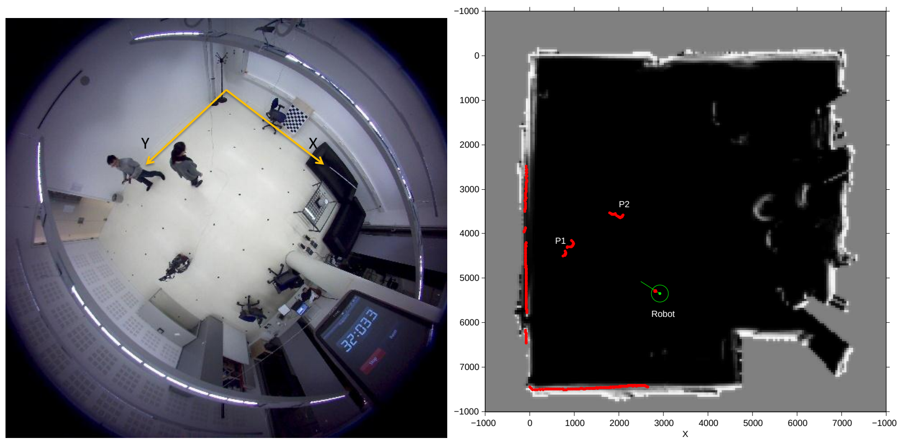

Fig. 3. An overview of the experiment room and the observed data. Left: captured by the over head camera; Right: laser detection points (red dots)

The Nomad robot runs on the Robot Operating System (ROS), and all data captured on the robot site is time stamped in ROS.

The exact time stamp of each frame collected with the overhead camera is obtained by means of a stopwatch mounted close to the camera. We use a nearest-neighbors classifier to recognize digits in the image to recover the time stamp. We synchronized the robot sensors and the overhead camera based on specific time points, where an event (e.g. the puncturing a balloon in front of the Laser Range Finder) was observed by both the robot sensor and the overhead camera.

The ground plane is subdivided into small cells of 50×50mm. In a first training run, the robot was remote-controlled to generate the background model. Second, the human model is trained according to Equ. 3. During testing, the two models are combined probabilistically into an occupancy map given the particles, as depicted in Fig. 4. Here each pixel of the occupancy map reflects the probability that that location is occupied, either by a person or by a background object in the room.

We evaluate the systems by measuring the Euclidean distance between the detection results and the ground truth locations of persons. In this paper, three localization approaches are tested and compared: a) with a single Laser Range Finder; b) with a single over head camera; c) with our proposed fusion framework. We evaluate the proposed system and the single modality approaches on 165 camera frames together with synchronized laser data. For each of these frames, volunteers manually annotated the locations of the persons in the ground plane, based on physical markers that were positioned on the floor during the recording, and these markers were used as reference to compute the ground truth location.

A particle sampling approach is applied both in the single laser and the data fusion approach. An equal number of 800 particles are sampled. Due to the fact that humans are not likely to be too close to each other, we define the safe distance between two persons as 500mm. We incorporate such assumption to reduce the space when sampling

Fig. 4. The occupancy map is generated from combining the human model and the background model. For each set of locations of persons, *i.e.* a particle, an occupancy map is estimated. Left: Human model. Middle: Background model. Right: Occupancy Map given the hypothetical location of persons (green crosses)

particles, *i.e.* the sampled point is always at least 500mm away from each of the points the previous sample set.

The single laser approach detects the foreground laser points by set a threshold to their probability in the background model. The threshold in our experiment is empirically set to 0.3. The particles are sampled from the foreground laser points with a Normal distribution on the location of the points. The weights are assigned by the likelihood of the laser data given the particles,and they are quantized in the sub-divided cells on the ground plane according to the locations of the particles. The human is then localized by recursively finding the cell that has the largest sum of weights as in [7].

In the approach with a single camera, we adopt the human detection algorithm from [7]. For each candidate location of the persons on the ground plane, the likelihood of the camera frame is measured. The locations of the multiple persons are found by choosing the locations that maximize the likelihood of the camera image.

The proposed approach combines the over head camera and the robot Laser Range Finder in a probabilistic Bayesian framework. After persons are localized with the single camera, the particles are sampled around the location of the persons with a Normal distribution. These particles are then weighted by the likelihood of the laser observations. The final detection is computed by the weighted sum of the particles that are sampled from the same person.

Fig. 5 shows the detection results of our data fusion system comparing with the approach using single modality. The proposed fusion system consistently outperforms the single-camera and the single-laser approach, and approximately 80 percent of the detections are less than 200 mm from the ground truth location. In contrast, only 70% of the camera-only detections and 27% of the laser-only detections are within such distance of the ground truth.

8 Conclusion and Future Work

We have proposed a novel probabilistic fusion framework for the localization of humans using ambient cameras and robot-mounted Laser Range Finders. Our experiments show

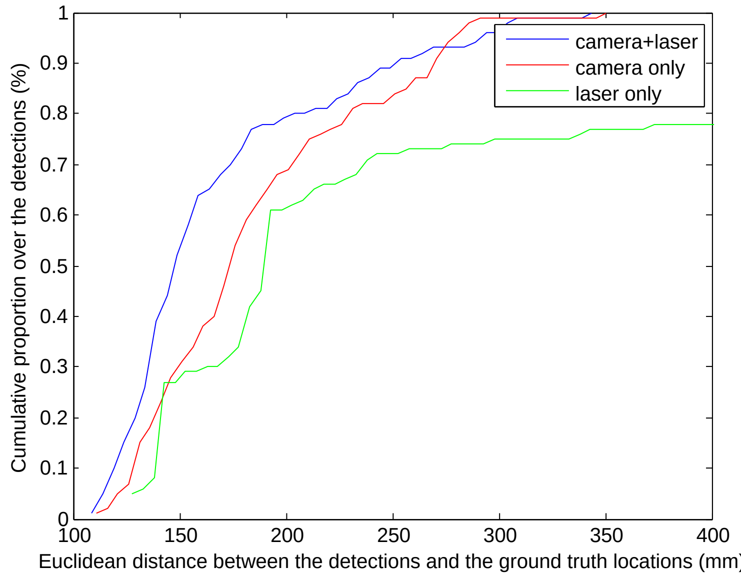

Fig. 5. Comparing the proposed data fusion approach and the single modality approach

substantial improvements in the accuracy of the localization, thus enabling more precise interaction between robot and humans. Due to its probabilistic nature, our framework can deal with occlusions and the absence of measurements in a principled way. As a result, the localization of humans is more robust, and natural interaction becomes possible even in challenging conditions.

In our current experimental work, the orientation and the location are not considered as part of the particle, but only the location of multiple persons are sampled. But we expect the performance can be improved by incorporating robot location and orientation into particles. We plan to specifically address occlusions and missed detections in one of the sensors. We will also extend the method to use more and different sensors, including the robot-mounted Kinect camera, as well as multiple overhead cameras.

Acknowledgment

This research is funded by the EU FP7-287624 Acceptable roboticCs COMPanions for AgeiNg Years (ACCOMPANY) and by the SIA project BALANCE-IT.

References

1. K.O. Arras, S. Grzonka, M. Luber, and W. Burgard. Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 1710–1715. IEEE, 2008.

2. N. Bellotto and H. Hu. Vision and laser data fusion for tracking people with a mobile robot. In *Robotics and Biomimetics, 2006. ROBIO'06. IEEE International Conference on*, pages 7–12. IEEE, 2006.
3. A.F. Bobick, S.S. Intille, J.W. Davis, F. Baird, C.S. Pinhanez, L.W. Campbell, Y.A. Ivanov, A. Schütte, and A. Wilson. The kidsroom: A perceptually-based interactive and immersive story environment. *Presence*, 8(4):369–393, 1999.
4. G. Bradski and A. Kaehler. *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, 2008.
5. Y. Cai, N. de Freitas, and J. Little. Robust visual tracking for multiple targets. *Computer Vision–ECCV 2006*, pages 107–118, 2006.
6. N. Checka, K.W. Wilson, M.R. Siracusa, and T. Darrell. Multiple person and speaker activity tracking with a particle filter. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 5, pages V–881. IEEE, 2004.
7. G. Englebienne and B.J.A. Kröse. Fast bayesian people detection. In *proceedings of the 22nd benelux AI conference (BNAIC 2010)*, 2010.
8. A. Fod, A. Howard, and MAJ Mataric. A laser-based people tracker. In *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*, volume 3, pages 3024–3029. IEEE, 2002.
9. B. Graf. Reactive navigation of an intelligent robotic walking aid. In *Robot and Human Interactive Communication, 2001. Proceedings. 10th IEEE International Workshop on*, pages 353–358. IEEE, 2001.
10. M. Kleinehagenbrock, S. Lang, J. Fritsch, F. Lomker, GA Fink, and G. Sagerer. Person tracking with a mobile robot based on multi-modal anchoring. In *Robot and Human Interactive Communication, 2002. Proceedings. 11th IEEE International Workshop on*, pages 423–429. IEEE, 2002.
11. B. Kluge, C. Kohler, and E. Prassler. Fast and robust tracking of multiple moving objects with a laser range finder. In *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, volume 2, pages 1683–1688. Ieee, 2001.
12. J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun. Towards robotic assistants in nursing homes: Challenges and results. *Robotics and Autonomous Systems*, 42(3):271–281, 2003.
13. O. Punska. Bayesian approaches to multi-sensor data fusion. *Cambridge University, Cambridge*, 1999.
14. C. Schlegel, J. Illmann, H. Jaberg, M. Schuster, and R. Wörz. Vision based person tracking with a mobile robot. In *British Machine Vision Conference*, pages 418–427, 1998.
15. X. Song, J. Cui, X. Wang, H. Zhao, and H. Zha. Tracking interacting targets with laser scanner via on-line supervised learning. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 2271–2276. IEEE, 2008.
16. K. Wada and T. Shibata. Living with seal robotsits sociopsychological and physiological influences on the elderly at a care house. *Robotics, IEEE Transactions on*, 23(5):972–980, 2007.
17. W. Zajdel, Z. Zivkovic, and BJA Krose. Keeping track of humans: Have i seen this person before? In *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pages 2081–2086. IEEE, 2005.
18. Z. Zivkovic and B. Krose. Part based people detection using 2d range data and images. In *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, pages 214–219. IEEE, 2007.