J Control Autom Electr Syst (2015) 26:616–629
DOI 10.1007/s40313-015-0202-6

A Probabilistic Approach for Fusing People Detectors

Natália C. Batista[1] · **Guilherme A. S. Pereira**[2]

Received: 6 April 2015 / Revised: 3 June 2015 / Accepted: 10 July 2015 / Published online: 31 July 2015

Abstract Automatic detection of people is essential for automated systems that interact with persons and perform complex tasks in an environment with humans. To detect people efficiently, in this article it is proposed the use of high-level information from several people detectors, which are combined using probabilistic techniques. The detectors rely on information from one or more sensors, such as cameras and laser rangefinders. The detectors' combination allows the prediction of the position of the persons inside the sensors' fields of view and, in some situations, outside them. Also, the fusion of the detector's output can make people detection more robust to failures and occlusions, yielding in more accurate and complete information than the one given by a single detector. The methodology presented in this paper is based on a recursive Bayes filter, whose prediction and update models are specified in function of the detectors used. Experiments were executed with a mobile robot that collects real data in a dynamic environment, which, in our methodology, is represented by a local semantic grid that combines three different people detectors. Results indicate the improvements brought by the approach in relation to a single detector alone.

This work was supported by Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG). Guilherme Pereira is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil.

✉ Natália C. Batista
nataliabatista@unifei.edu.br

Guilherme A. S. Pereira
gpereira@ufmg.br

[1] Federal University of Itajubá - UNIFEI, Rua Irmã Ivone Drummond 200, Itabira, MG 35903-087, Brazil

[2] Federal University of Minas Gerais - UFMG, Av. Antônio Carlos 6627, Belo Horizonte, MG 31270-901, Brazil

Keywords People detection · Information fusion · Bayes filter · Semantic grid

1 Introduction

In the next years, the development of robotic systems will dramatically change the way humans move, work, or have fun. The coexistence of such systems and human beings is increasing everyday, and because of that, robots should be designed to interact safely with people. To allow this interaction, the robots should efficiently detect the people in their workspace.

Robots may execute complex tasks to help people or cooperate with them, for instance by doing home cleaning, guiding newcomers in a museum, or carrying a load (Pereira et al. 2013). As defined by the International Federation of Robotics (IFR), robots that perform such tasks are called service robots. A critical issue for service robots in an environment with people is human–robot interaction. Robots need to be aware of the presence of people, and in some cases, it is desirable that the robot is controlled or supervised by a human operator (Ceccarelli 2011). Therefore, it is essential that the robots have an advanced perception system, which is responsible for transforming raw sensor data, such as camera images and laser scans, into consistent and useful information not only to understand the environment and detect objects, but also to detect people and their location. In this context, autonomous vehicles are examples of service robots with the need for people detection (Geronimo et al. 2010). People detection is also important in smart environments where it can be used to predict the behavior of the users (Hofmann et al. 2011).

Robotic systems for people detection usually seek for candidates in the field of view (FOV) of the sensors using characteristics such as shape, symmetry, texture, movement,

and frequency of motion of human legs (Broggi et al. 2009). The most common sensors used in this area are laser rangefinders, radars, and cameras.

Laser rangefinders and radars are distance sensors that scan one or more planes in space, obtaining the distance from obstacles in relation to its reference frame. The approaches that detect people with such sensors generally perform the extraction of geometric features such as border size, convexity, number of points, lines, and corners. The features are used to train classifiers or to calculate thresholds (Premebida et al. 2009; Spinello and Siegwart 2008). Other approaches are based on pattern matching, for example in the works by Pereira et al. (2013), Oliveira et al. (2010), and Bellotto and Hu (2009). The approaches which are not based on features apply local minimum search, detection based on motion or background subtraction (Cui et al. 2005).

People detectors in camera images are often based on sliding window, segmentation, or keypoint approaches (Dollar et al. 2012; Varga et al. 2014). The computation of Haar features in the image and the use of a cascade structure for detection with AdaBoost feature selection is an approach that serves as a foundation for modern detectors (Pereira et al. 2013; Bellotto and Hu 2009). Another popular technique for people detection in images is the histogram of oriented gradient (HOG), introduced by Dalal and Triggs (2005). The authors perform people and object recognition even in complex environments and under variable lighting conditions using HOG descriptors classified by support vector machines (SVM). The basic idea is that the local appearance and shape of objects can be characterized by the local distribution of intensity gradients, which represent the direction of the edges. These descriptors are used in several works such as (Varvadoukas et al. 2012; Oliveira et al. 2010), and (Spinello and Siegwart 2008). State-of-the-art detectors rely on sliding windows over feature pyramids, which are multi-scale representations of an image with fast construction, allowing real-time performance such as some extensions of the work by Dollar et al. (2014), which uses aggregated features (normalized gradient magnitude, HOG, and LUV color channels) and AdaBoost.

Data from laser, radar, and camera can be used simultaneously in order to combine the technological advantages of each sensor and to compensate for their limitations. The combination of sensing information, known as sensor fusion or sensor integration, is frequently used in several robotic applications, such as autonomous navigation, object classification, and localization (Antunes et al. 2012). In these cases, the combination may overcome problems caused by occlusion, direct sunlight, low reflectivity of dark objects, bad weather conditions, among others.

For people detection, there are approaches which combine sensor data at feature level or at classifier (or object) level. In the feature level, raw data from sensors are processed to obtain features, such as lines, corners, height, speed, and motion, which are fused to be classified as people or not people later. The works by Cho et al. (2014), Utasi and Benedek (2013), and Bota and Nedesvchi (2008) are based on this approach. The second approach, which is used in the works by Premebida et al. (2014), Huerta et al. (2014), and Oliveira et al. (2010), performs people or object detection on sensor data separately and then combines the candidates detected. The methodology presented in this paper lies in this second approach.

Among the methodologies found in the literature for people detection, several can be classified as Bayesian approaches. Either they directly use the Bayes' rule or they are based on the various techniques derived from the Bayes' filter (Kalman filter, particle filters, occupancy grid, etc). The Bayes' rule is generally used to, given the extracted characteristics and some confidence about them, compute the probability that they represent a person. This computation may or may not use information from the past to compute the current probability. Examples of works that use such an approach are (Utasi and Benedek 2013; Bota and Nedesvchi 2008; Ngako Pangop et al. 2008). On the other hand, when the variations of the Bayes' filter are used, the method includes a prediction step, based on mathematical models and information from the past, and a correction step, when sensor data are used to update the prediction. Some works that follow this idea are (Cho et al. 2014; Gidel et al. 2009; Monteiro et al. 2006).

The methodology proposed in this work can be classified as a Bayesian approach, falling in the subclass of works that uses variations of the Bayes' filter. In our method, the Bayes' filter is used to combine people detectors. The information to be combined is data already processed and classified as people by a set of previously published classifiers. The combined information has a level of confidence that is higher than the one from each individual people detector. This information is available to the applications (people tracking, robot navigation, human–robot interaction, etc) as a set of numbers that indicate the probability that specific regions of the environment are occupied by people. A block diagram of our solution is presented in Fig. 1. The main contributions of the proposed methodology in relation to other Bayesian approaches for people detection are: (1) A prediction step executed in two phases: one based on a sensor motion model and the other based on a people motion model; (2) a new people motion model, which considers the probability distribution of people in regions of the workspace and their probability of movement; and (3) the use of high-level information from multiple people detectors in the correction step using models based on precision and false-negative rate for these detectors. These contributions indicate that, even being based on a mature theory, the methodology proposed in this work constitutes a novel strategy for people detection.

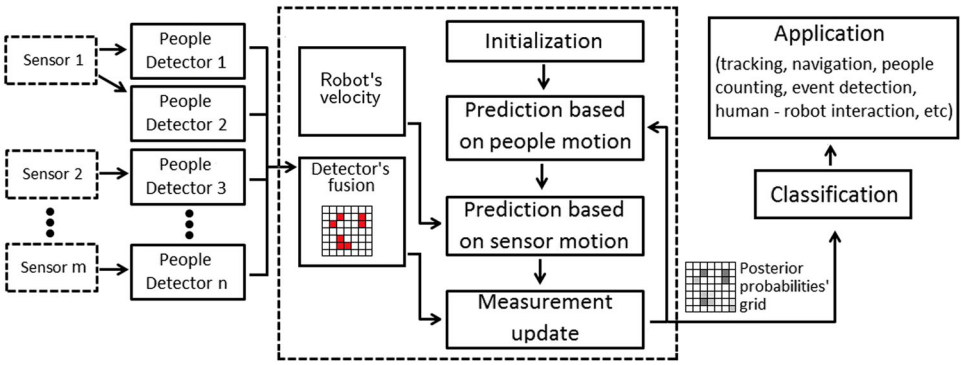


Fig. 1 Overview of the steps of the proposed approach, which is enclosed by the *biggest dashed rectangle*. In addition, the scheme shows other steps to integrate the approach to some application

Although the proposed approach can be implemented using several techniques, including efficient Monte Carlo variations, such as the particle filter, in this paper it was implemented and tested using a semantic occupancy grid. Occupancy grids are being increasingly employed in robotics applications (Liu and von Wichert 2014; Adarve et al. 2012) for being a compact representation of the environment close to the robot. It also provides information on the occupancy of the space, allowing the representation of other detected objects, and reduces the problem of data association from multiple sensors to a simple mapping to a cell of the grid (Yoder et al. 2010). In the experiments presented in this work, an occupancy grid was built to represent the local workspace of a mobile service robot navigating in a building with people inside. The robot is equipped with a laser rangefinder and a monocular camera and moves autonomously through the environment, which is subjected to large illumination changes. The results of this experiment are quantitatively and qualitatively analyzed, showing an increase in the hit rate of the detections and in the accuracy of people location when compared with the results from single detectors.

In the next section, we briefly introduce the main theoretical concepts necessary to understand the proposed theory and its implementation using semantic grids.

2 Background

The objective of the Bayesian approach proposed in this paper is to determine the chances that a given region of the space is occupied by a person. Thus, given a 2D workspace partitioned into regions, we want to determine the probability of having a person in a region centered at point (x, y), assuming that we have a probabilistic model for the motion of the people, a set of stochastic information given by an ensemble of sensor-based people detectors, and a probabilistic model for the motion of the sensors in the workspace.

Let X be the random variable that represents the state of the region centered on (x, y) in relation to the type of object contained in that region. If the space under consideration is, for example, an area in the road in front of a vehicle, the experiment to observe this region at time t may result in the conclusion that there is a person, a car, a bicycle, an animal, other obstacle, or nothing (the area is free). In this case, the random variable X can assume the following values: person, car, bicycle, animal, other obstacles, or free space. Therefore, the sample space S consists of the possible outcomes obtained from observation, such that $S = \{\text{person, car, bicycle, animal, } \ldots, \text{ free space}\}$. According to the second axiom of probability (Papoulis and Pillai 2002), the sum of the probabilities of all events in the sample space is equal to one:

$$\sum_{x \in S} P(X = x) = P(X = \text{person}) + P(X = \text{car}) + \cdots + P(X = \text{free space}) = 1. \quad (1)$$

Since the focus of this work is the detection of people, we consider that the sample space is partitioned into two subsets: $S = \text{p} \cup \text{np}$, where p is the event $\{X = person\}$ and np (not person) is the event in which can occur other elements of S, except for person. In this paper, for the sake of simplicity in the notation, the probability that event p occurs or $P(X = \text{p})$ will be denoted by $P(\text{p})$. The computation of $P(\text{p})$ depends on several factors that should be considered, such as: the state of the region at the time of the last observation, the displacement of people from other neighboring regions to the region of interest, the information from the sensors that observe the region, and the movement of these sensors. The Bayes filter, introduced next, is applied to consistently consider all these factors.

2.1 The Bayes Filter

Consider X to be a random variable and x a specific value that X can assume. $P(x)$ is the probability that X becomes x. The Bayes Theorem is based on the theorem of total probability and on the conditional probability rule (Papoulis and Pillai 2002). It relates a conditional probability of the type $P(x|y)$ to a conditional probability $P(y|x)$ as:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}. \quad (2)$$

In the field of probabilistic robotics, the Bayes theorem is used to infer a quantity x from data y (e.g., a sensor measurement). The quantity x is a state that can be, for example, the pose or the velocity of a robot, its localization, and/or features and objects in the environment. The distribution $P(x)$ is known as *p*rior probability distribution (before the experiment is conducted) which represents knowledge about x before incorporating y. The probability $P(x|y)$ is called the *p*osterior probability distribution (after the experiment is conducted) (Thrun et al. 2005).

The Bayes filter is derived from the application of the Bayes theorem to the posterior probability $P(x_t|z_{1:t}, u_{1:t})$, which determines the state x conditioned on the sensor measurements $(z_{1:t})$ and on information about the change in state in the environment, which is called control data $(u_{1:t})$. As states may change over time and measurement and control are performed at certain time instants, the subscript values of the random variables indicate the time instant considered in discrete values. For example, $z_{1:t}$ represents the set of all measurements acquired from the time instant equal to 1 until time t. The notation used here is similar to the one in (Thrun et al. 2005), which describes the Bayes filter as a way to compute the posterior probability, $P(x_t|z_{1:t}, u_{1:t})$, denoted by $bel(x_t)$, in two steps:

$$\overline{bel}(x_t) = \int P(x_t|u_t, x_{t-1})bel(x_{t-1})\mathrm{d}x_{t-1} \quad (3)$$

$$bel(x_t) = \eta P(z_t|x_t)\overline{bel}(x_t). \quad (4)$$

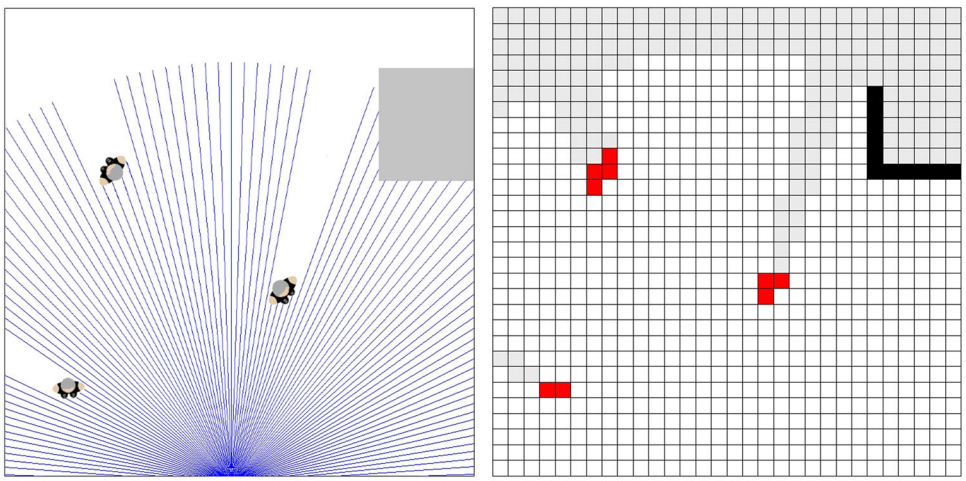
The Bayes filter is recursive as it calculates the belief distribution $bel(x_t)$ at time t by using the distribution $bel(x_{t-1})$ at time $t-1$. Equation (3) takes into account control u_t and the distribution of the state x_{t-1} to calculate the distribution of the state x_t. This step is known as prediction. When the state space is finite, the integral in (3) becomes a finite sum. Equation (4) is known as measurement update and takes into account the measurement z_t. Constant η is used for normalization, ensuring that the resulting product is a probability function and its integral is equal to 1.

In practical applications, Gaussian filters (such as the Kalman Filter and its variations) and nonparametric filters are tractable implementations of the Bayes filter for continuous spaces. Nonparametric approaches approximate posteriors by a finite number of values, each roughly corresponding to a region in the state space, such as particle filters, histogram filters, and occupancy mapping algorithms (Thrun et al. 2005). The implementation presented in this work uses semantic grids, which are based on occupancy grids. Semantic grids are presented in next subsection.

2.2 Semantic Grid

An occupancy grid is a stochastic representation of spatial information of the environment (Elfes 1990). It is arranged in cells of the same size, each associated with a (x, y) coordinate and a probability of occupancy. As defined in (Thrun et al. 2005), $\mathbf{m}_i$ is the cell with index i. The occupancy grid, m, partitions the space into a finite number, N, of cells as $m = \{\mathbf{m}_i | 1 \leq i \leq N\}$. Each cell $\mathbf{m}_i$ has a value of occupancy, usually binary, indicating whether the cell is occupied or free. The probability of a cell being occupied is referred to as $P(\mathbf{m}_i)$.

A semantic grid is an occupancy grid that integrates the spatial representation of the environment with the poses of objects of known classes (Nüchter and Hertzberg 2008). In this work, the semantic grid will provide information about the presence of people in the environment and a measure of the confidence of this information. An example of semantic grid is in Fig. 2. Figure 2a shows a scene containing three



(a) **(b)**

Fig. 2 **a** A scene containing people and another object being scanned by a laser positioned at the middle bottom. **b** A semantic grid that represents the scene colored according to the information obtained by the sensor's readings

people and an object being scanned by a laser rangefinder positioned at the middle bottom (its rays are the blue lines) in a bird's eye view. Figure 2b shows a semantic grid with 900 (30×30) cells that represents the scene in (a). The grid cells are colored according to the readings of the laser: white for free cells, red to cells occupied by people, black to cells occupied by obstacles and gray to cells outside the sensor range.

Given the definitions of this section, next section presents our methodology to combine information from multiple people detectors using the Bayes filter. This filter is experimentally tested using a semantic grid, as shown in Sect. 4.

3 Metodology

The proposed approach applies the Bayes filter to (i) predict the presence of people in a specific region of the space and (ii) correct this first estimate based on the information given by an ensemble of people detectors. The following subsections detail each step of the approach.

3.1 Prediction

In this work, we propose the use of two prediction phases: one based on people motion and another based on sensor motion. These phases are detailed next.

3.1.1 Prediction Based on People Motion

The prediction based on people motion is a step that computes the probability of a particular region to be occupied by people based on knowledge on the people motion and on knowledge of the probability of existence of people, in this and in the neighboring regions, at the preceding time $(t-1)$. Therefore, given the probability $bel(x_{t-1})$ of the presence of people at (x, y) in the preceding time $(t-1)$ and the people motion model, represented by a probabilistic density function (PDF), the probability that the region centered at (x, y) is occupied at time t is computed by Eq. (5) which is based on the prediction equation of the Bayes filter.

$$\overline{bel}'(x_t) = \int P(x_t|x_{t-1}, v_{1:t}^{\text{people}}) bel(x_{t-1})\, \mathrm{d}x_{t-1}\,. \tag{5}$$

Since state x can only assume two values, $x = \text{p}$ or $x = \text{np}$, the state space is discrete and the integral in (5) becomes a finite sum. At time t, the region centered at (x, y) will be occupied by people in two situations: if there were any people at (x, y) at time $t-1$ and they stood still, or if there were no people at time $t-1$ but some people moved from a second region (i, j) to (x, y). Thus, Eq. (5) becomes similar to the equation used to compute the probability of occupancy of regions by species in the field of ecology, which also takes into account the movement of individuals (MacKenzie et al. 2003). This model considers the occupation of areas (colonization) and its unemployment (extinction), but with different speed and probabilities of colonization and extinction. By rewriting Eq. (5) based on this model and assuming that the state is complete, that is, knowledge of x_{t-1} imply that past measurements and information about past velocities do not contribute to determining the state x_t, we obtain:

$$\begin{aligned}\overline{bel}'(x_t = \text{p}) = {} & P(\text{p}|x_{t-1} = \text{p}, v_t^{\text{people}} = 0) bel(x_{t-1} = \text{p}) \\ & + P(\text{p}|x_{t-1} = \text{np}, v_t^{\text{people}} = v_{\text{average}}) bel(x_{t-1} = \text{np})\,,\end{aligned} \tag{6}$$

where:

- $\overline{bel}'(x_t) = P(x_t|z_{1:t-1}, v_{1:t}^{\text{people}})$ is the probability of state x_t at time t conditioned on all past measurements $z_{1:t-1}$ and people velocities $v_{1:t}^{\text{people}}$.
- $bel(x_{t-1}) = P(x_{t-1}|z_{1:t-1}, v_{1:t-1}^{\text{people}})$ represents the prior belief over state x_{t-1}, i. e., the probability of the state x_{t-1} conditioned to all past sensor measurements $z_{1:t-1}$ and past people velocities $v_{1:t-1}^{\text{people}}$.
- $P(\text{p}|x_{t-1} = \text{p}, v_t^{\text{people}} = 0)$ is the probability of people do not move way from their current region (x, y).
- $P(\text{p}|x_{t-1} = \text{np}, v_t^{\text{people}} = v_{\text{average}})$ is the probability of people come to occupy region (x, y) given that their velocity is v_{average}.

The probability $P(\text{p}|x_{t-1} = \text{p}, v_t^{\text{people}} = 0)$ is given by:

$$P(\text{p}|x_{t-1} = \text{p}, v_t^{\text{people}} = 0) = p_{\text{stationary}}\,, \tag{7}$$

where $p_{\text{stationary}}$ is the probability of people to remain in their current position. On the other hand, probability $P(\text{p}|x_{t-1} = \text{np}, v_t^{\text{people}} = v_{\text{average}})$ is computed as:

$$\begin{aligned}P(\text{p}|x_{t-1} = \text{np}, v_t^{\text{people}} = v_{\text{average}}) = \\ = 1 - \frac{p_{\text{free}}}{p_{\text{free}} + \sum_{k=(i,j)\in G} \frac{p_{\text{free}}\, p_k\, bel(x_{t-1}^{i,j}=\text{p})}{(1-p_k\, bel(x_{t-1}^{i,j}=\text{p}))}}\end{aligned} \tag{8}$$

where p_{free} is the probability of the region (x, y) to remain free (without people) and p_k is the probability of people to move from (i, j) to (x, y), which can be computed by a people motion model. These probabilities are related by:

$$p_{\text{free}} = \prod_{k=(i,j)\in G} 1 - p_k\,. \tag{9}$$

The set G in Eqs. (8) and (9) is the space of all regions that may be occupied by people. In this work, we are considering

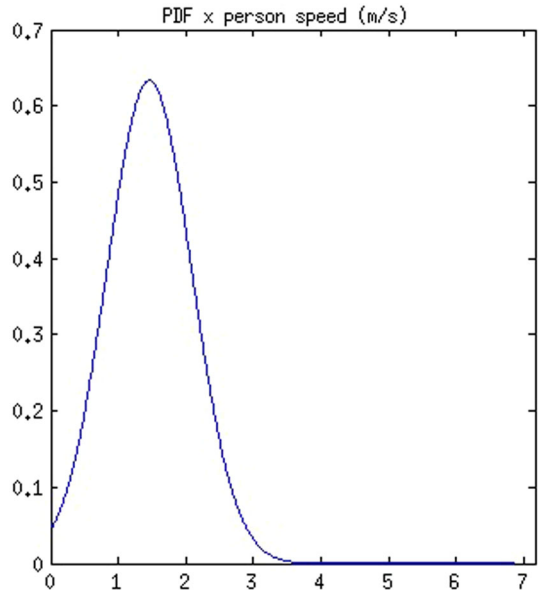


Fig. 3 People motion model: probability density function (PDF) × speed (m/s)

that this space is discrete. Moreover, in practical implementations, G can be restricted to the neighborhood determined by all locations that a person can reach from (x, y) in a single time interval.

The people motion model describes the distribution of velocities at which people usually walks. There is a physical limit to the maximum speed attainable by a person. Although unlike for ordinary people, in this work the limit to this speed is the average speed reached by the fastest athlete in the world (Hogenboom 2013). Based on this, the people motion model considers a low probability to high-speed movements and a higher probability to speeds around the average speed with which people usually move. The literature suggests that the people motion model can be represented by a normal PDF with mean value of 1.46 m/s and standard deviation of 0.63 (Daamen and Hoogendoorn 2007). A plot of this model is in Fig. 3.

It is also necessary to include in the previous model the probability that people may stay in the same region between two periods of time. Assuming that the probability of a person to remain standing still in urban spaces is $p_{stationary}$, we propose the following model:

$$p_k = \begin{cases} p_{\text{stationary}}, \text{ if } dist((x, y), (i, j)) = 0; \\ (1 - p_{\text{stationary}}) \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(dist((x,y),(i,j))/\Delta t - \mu)^2}{2\times\sigma^2}}, \text{ if} \\ dist((x, y), (i, j)) \neq 0. \end{cases}$$

where Δt is the time interval between t and $t - 1$ and function $dist(\cdot, \cdot)$ computes the Euclidean distance between two points in a two dimensional space.

In this model, it is assumed that people can walk everywhere (no knowledge about obstacles or the environment map) and in every direction. Thus, as the velocity has two components (linear and angular velocity), the linear component has a distribution function equivalent to the distribution of speed and the distribution of angular velocity is uniform (it is considered that people can move to any direction with the same probability).

The second step of the prediction is detailed in the next subsection.

3.1.2 Prediction Based on Sensor Motion

In the second step of the prediction, the probability of the existence of people in a determined region of the scene relative to the sensor frame is computed according to the displacement of the sensors. Assuming that the sensors are attached to a mobile robot, the motion model can rely on information given by the robot's sensors, such as odometers, inertial measurement units or GPS. Since this information is subject to errors, a distribution function of the sensors' position at time t indicates that it is possible that the sensors are in several other positions with some probability. Consequently, the estimate of the people's position relative to the sensors' frame is subject to the same uncertainties. To calculate the change in the coordinate system, i.e., the new position of the people after sensors' motion, the displacement vector of the sensor can be subtracted from the people's coordinates. Thus, even though none of the people in the environment has moved, relative motion can occur when the sensor moves.

Considering that the prediction based on people's motion is computed using (6), the influence of the sensor motion is then given by:

$$\overline{bel}(x_t = \mathrm{p}) = P(\mathrm{p}|v_{1:t}^{\text{sensors}})\overline{bel}'(x_t = \mathrm{p}), \tag{10}$$

which can be written as:

$$\overline{bel}(x_t = \mathrm{p}) = P(\mathrm{p}|z_{1:t-1}, v_{1:t}^{\text{people}}, v_{1:t}^{\text{sensors}}). \tag{11}$$

In practical situations, the actual velocity of the robot differs from the measured velocity, which can be approximated by a random variable. The probability $P(\mathrm{p}|v_{1:t}^{\text{sensors}})$ can be found using Eq. (8), when v_t^{people} is replaced by v_t^{sensors}. The velocity motion model described by Thrun et al. (2005), with a normal distribution is used to calculate the probability p_k.

Next subsection describes the update step of the methodology.

3.2 Measurement Update

The measurement update is a step that follows the prediction and involves determining the probability that a given region

of space is occupied by people using information from the sensor-based people detectors and the probability computed in the prediction step. This step, based on the second equation of the Bayes filter (Eq. (4)), is also known as correction, since it incorporates a new measurement z_t to the belief by multiplying $\overline{bel}(x_t)$ by the probability of the measurement z_t has been observed:

$$bel(x_t = \mathrm{p}) = \eta P(D_1, \ldots, D_N | x_t = \mathrm{p}) \overline{bel}(x_t = \mathrm{p}) , \quad (12)$$

where the normalization step is given by $\eta = 1/(bel(x_t = \mathrm{p}) + bel(x_t = \mathrm{np}))$ and $P(D_1, \ldots, D_N | x_t = \mathrm{p})$ is the probability of detectors $D_1, D_2 \ldots, D_N$ indicate the presence of people in a region when there is in fact people at the region. This term is computed as explained below.

The proposed approach uses information from multiple people detectors in the update step. Thus, the raw information from the sensors is processed so that each detector provides the position of the detected people and a measure of the confidence for such estimations. This high-level information is combined so that when more than one detector indicates a person at a given position, the detection's confidence is greater than the one relative to the cases when only one detector detects a particular person.

When more than one detector is used, their information can be combined by taking into account their different characteristics. To combine multiple sensors, there are many possibilities in the sensor fusion literature. Among them, we find linear opinion pools (Adarve et al. 2012; Baig et al. 2014), Bayesian fusion (Yguel et al. 2006), and maximum functions (Thrun et al. 2005). As described in Baig et al. (2014), conflicting information may generate errors in some of these methods. To deal with conflicting information and to shorten the confidence of those detectors that do not give relevant information to the process, in this work, the information from all detectors of different sensors are merged based on De Morgan's law (Thrun et al. 2005). Considering that there are N detectors $D_1, D_2, \ldots, D_N$, the confidence of the fusion when at least one of the detectors detects people in a given region of the space is written as:

$$P(D_1, \ldots, D_N | \mathrm{p}) = 1 - \prod_{i=1:N} (1 - P(D_i | \mathrm{p})) . \quad (13)$$

To compute $P(D_i|\mathrm{p})$, positive predictive value (PPV) and false-negative rate (FNR) were used. PPV, also known as precision, is the ratio between the number of correct detections and the total number of detections. It measures the number of hits in relation to the total number of objects identified by the classifier as positive. In our case, it is used when the detector detects people. Thus, in these cases $P(D_i|\mathrm{p}) = \mathrm{PPV}$. When the detector does not detect people $P(D_i|\mathrm{p}) = \mathrm{FNR}$, which is the ratio between the number of false negatives and total number of objects actually in the scene. Values of PPV and FNR are found experimentally for each detector.

When none of the detectors detects people, De Morgan's law is not used. To ensure that the probability of people in the region is reduced, the confidence associated with the presence of people in this region will be the smaller FNR of the detectors:

$$P(D_1, \ldots, D_N | \mathrm{p}) = \min_i \ FNR_{D_i} . \quad (14)$$

In our experiments, shown in the next section, three detectors were used to validate our methodology. Details of the experiments are shown next.

4 Experimental Results

In this section, we show some of the characteristics of our methodology using real-world data. Our experimental setup consists of a laser rangefinder and a camera fixed on a mobile robotic platform called MARIA (manipulator robot for interaction and assistance). A picture of the robot and its sensors is shown in Fig. 4. Laser and camera were positioned 0.29 and 1.63 m above the ground, respectively. The laser was approximately aligned with the camera x-axis but shifted about 0.20 m ahead. The orientation of the camera relative to the laser was supplied by a MATLAB-based camera-laser calibration toolbox (Kassir and Peynot 2010). The orientation angles about axis x, y, and z obtained using this package were, respectively, 0.1222, 0.0034, and 0.0175 radians. The sensors' positions are in consonance with the height and orientation expected by the detectors used in this work. The laser

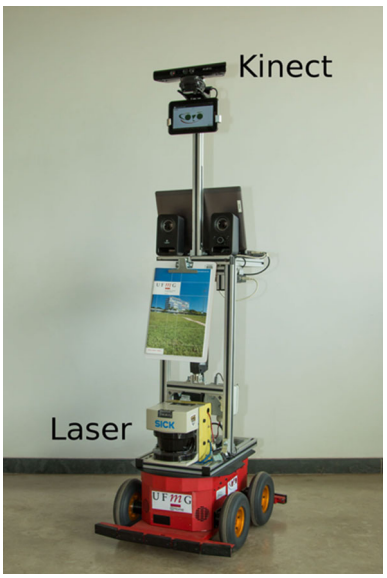


Fig. 4 Robot used in the experiments of this work

sensor used was the SICK LMS291-S05. It has a field of view (FOV) of 180° and was configured to the maximum range of 32 m with 1° of angular resolution. The camera used was the one from the Kinect sensor installed on the robot. Although the Kinect can provide depth information, in our experiments we only used the images generated by the RGB camera. The size of each image is 640 × 480 pixels. The horizontal FOV of the camera is of 62°.

In our experiments, the robot navigates autonomously in an office-like environment using the methodology proposed by Araujo et al. (2015). The environment considered in this work is a building that has open and closed corridors, which causes significant changes in the illumination conditions. Although these changes turn the environment similar to outdoors environments in terms of illumination, it is important to emphasize that the environment is still an indoor environment in the sense that it presents an even and flat ground, which does not cause any difficult to the data acquisition process. During the navigation, the robot speed ranged from 0 to 0.6 m/s. To collect data from laser and camera simultaneously, ROS (robot operating system) was used.[1] The original frequency of the camera was 30 Hz, but it was down-sampled to 10 Hz for faster processing. The laser frequency was 9 Hz. As the frequencies of the sensors differ, synchronization was performed by taking the laser readings temporally closer to the last image of the camera.

After laser and camera data were collected, three different people detectors were used: two of them based on laser (Detector 1 by Bellotto and Hu (2009) and Detector 2 by Spinello and Siegwart (2008)) and the other one based on images (Detector 3 by Dollar et al. (2014)). Although these detectors are not state-of-the-art detectors, they were chosen because they are freely available and have presented satisfactory results in previously published works (Pereira et al. 2013; Varvadoukas et al. 2012; Benenson et al. 2015). Notice, however, that the proposed approach allows the use of any other people detector and different sensor configurations.

We have implemented our methodology using a local semantic grid of 30 × 30 square cells, each with a side measuring 0.3 m. Observe that the dimension chosen for the cells allows a person to occupy more than one cell, as well as two or more people share parts of the same cell. In our implementation, the constants of the equations in Sect. 3 are shown in Table 1. The values of PPV and FNR for the detectors were obtained experimentally using a set of data similar to the one used in this section.

The semantic grid is in the same plane as the laser in a situation similar to the one illustrated in Fig. 2a. The detectors provide the relative positions of the detected people, which are mapped to the local grid. As the sensors have limitations and are noisy, they have a degree of uncertainty regarding both, the location provided and the result of detection. The uncertainty in the position provided by the laser was discarded for being negligible compared to the dimensions of a person (laser position error is about 35 mm according to SICK (2006)). Regarding the image-based detector, since it used a single camera, the method proposed by Stein et al. (2003) to estimate the distance between the camera and the detected person was used. This is based on an approach that assumes the knowledge of the height of the camera in relation to the ground. The uncertainty of this estimation was obtained empirically. The PDF that represents the error of the estimation was approximated by a normal distribution with 0 mean and standard deviation of 1.0 m on x and 1.7 m on y. Figure 5 shows the effect of the distribution on the location of people detected by the camera in the cells (15, 15) and (15, 16) of the grid.

Table 1 Constants used in the experiments

Constant	Value
μ	1.46 m/s
σ	0.63 m/s
$p_{\text{stationary}}$	0.85
Δt	0.1 s
PPV (Detector 1)	0.95
FNR (Detector 1)	0.30
PPV (Detector 2)	0.50
FNR (Detector 2)	0.40
PPV (Detector 3)	0.95
FNR (Detector 3)	0.50

The proposed approach was prototyped in MATLAB. In an Intel Core2 Duo 1.8 GHz, the execution time of each iteration for a grid with 30 × 30 was about 1.78 seconds. To implement the method in real time, a C implementation and a parallelization of some steps of the approach using GPU should be considered. The use of GPU in other approaches has showed that it is possible to achieve a considerable decrease in processing time. In the work by Yoder et al. (2010), for example, the reduction in time was about 95 % by exploring the parallel structure of their method, which was also based on occupancy grids. It is important to notice that the computational complexity of the method is linear in relation to the number of cells in the occupancy grid.

Figure 6 shows two consecutive snapshots of our experiment. The first one is represented by figures (a) and (c)–(j) and the second one by figures (b) and (k)–(r). Figure 6a, b shows images of the camera with detected people marked by rectangles. The grid in figures (c) and (k) represents the ground truth data which were annotated manually using raw laser data projected on the grid. In this grid, only people were annotated and they are represented by red cells. Notice

[1] ROS is a software which provides libraries and tools for applications in robotics (http://www.ros.org/).

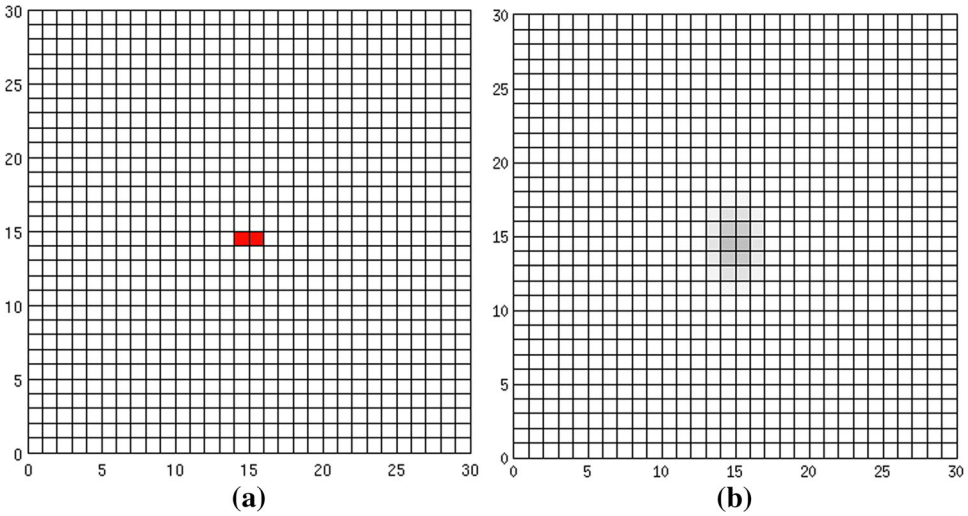


Fig. 5 Effect of the distribution on the location of people. **a** *Red cells* represent people as found by the people detector. **b** Localization of people detected in (**a**) after application of error distribution. In this figure, the *darker* is the cell, the greater is the probability of people (Color figure online)

in these figures that there are two persons in the scene: one of them positioned at the top left (Person 1) that occupies four cells and the other one in the bottom right (Person 2) that occupies two cells in figure (c) and three cells in (k). Person 1 and Person 2 are moving in opposite directions. The camera's field of view is smaller than the laser's, so the images in (a) and (b) do not show Person 2. Notice in figures (a) and (b) that there are some other people detected by Detector 3 that are not represented on the local grid, once they are over 9 m from the robot and thus beyond the grid limits.

Figure 6d, l show the prediction step based on people motion. In these figures, the darker is the cell, the greater is the probability of occupation by people. The prediction step is based on the belief computed in the previous time interval (prior). Therefore, (l) is computed by the application of the people's motion model to the grid in Fig. 6j, which is the final result of the method in the first iteration shown in Fig. 6. The prior related to Fig. 6d is not shown. By comparing (j) and (l) one can notice that the motion model spreads the probability around the grid. The prediction step considers that the probability of existence of people outside the grid is 0.5, meaning that there is no knowledge about the presence of people in this region. This forces the motion model to take into account that people from outside can move inside the grid.

Figure 6e, m shows the results of the prediction based on sensor motion. Since the robot was moving with linear velocity of 0.35 m/s and angular velocity of 0 rad/s, it is almost not possible to notice a small shift of the cells from up to down due to the motion of the robot. This shift would certainly be visible at higher velocities.

The grids in Fig. 6f–h, n–p show the results given by the detectors mapped into semantic grids. In these grids, the red cells correspond to people. Figure 6f refers to Detector 1, a laser-based detector which detected only Person 2 in the first time step. In the next iteration, this detector found only Person 1 (Fig. 6n). The other laser-based detector, Detector 2, in Fig. 6g found Person 1 but incorrectly detected another person at the bottom right of the grid. In the next instant of time, this detector again found Person 1 and another person, which was not a correct detection (Fig. 6o). Figure 6h, p shows the results of Detector 3, a vision-based detector. It successfully detected Person 1. Person 2 can never be detected using images, since it is outside the field of view of the camera. In Fig. 6h, p, the detection is represented by several cells because there is a large uncertainty in relation to the position of people detected with a single camera.

Fusion of the three detectors using Eqs. (13) and (14) is shown in Fig. 6i, q. Again, darker cells indicates a large probability of people. The free cells have a lower probability, which is represented by light gray. In the first time step, Person 1 and Person 2 were detected at least by one of the detectors and then, fusion shows a high probability in the cells occupied by them (Fig. 6i). On the other hand, notice in the fusion results of the second time step, shown in Fig. 6q, that only the position of Person 1 has a high probability, since Person 2 was not detected in this time step. Also, thanks to the low confidence assigned to Detector 2, the false positives (i.e., incorrect detections) in both time steps had their probabilities minimized during the fusion step.

The posterior probability, which was computed by the combination of the belief in the prediction step and the probability obtained by the fusion of the detectors as in Eq. (12), is shown in Fig. 6j, r. In Fig. 6j, the cells occupied by people have the largest probabilities, representing the most likely location of people. It shows that when two different people detectors detect the same person, the probability in the cells regarding this detection is enhanced. Figure 6r is the resulting grid after the update step in the second instant of time. The location of Person 1 (top left) is improved by increasing

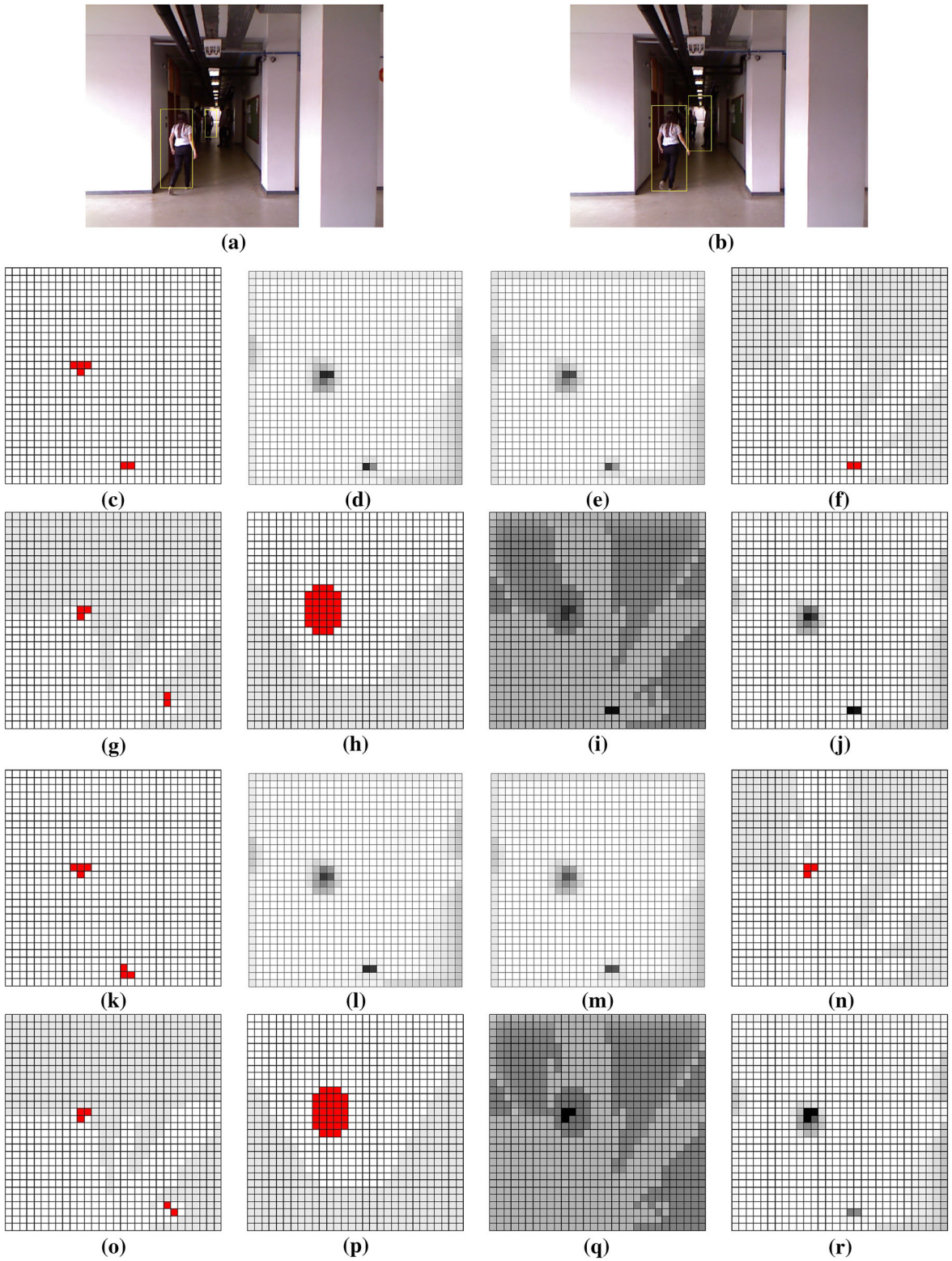


Fig. 6 Experimental results in two instants of time: Time t ((**a**) and (**c**)–(**j**)); time $t+1$ ((**b**) and (**k**)–(**r**)). The figures show: a camera image with detections ((**a**) and (**b**)); the ground truth with people in *red* ((**c**) and (**k**)); the predicted grid based on people motion ((**d**) and (**l**)); the predicted grid based on sensor motion ((e) and (m)); the detections of Detector 1 using laser ((**f**) and (**n**)); the detections of Detector 2 using laser ((**g**) and (**o**)); the detections of Detector 3 using image ((**h**) and (**p**)); the fusion of the detectors ((**i**) and (**q**)); and the final grid ((**j**) and (**r**)) (Color figure online)

the probability in the cells in which the laser-based detectors founds that person. Even though the three detectors failed to detect Person 2, the cells next to his actual location have a prominent probability (but a bit lighter than in Fig. 6j) because in the update step of the methodology the information from the prediction step, which is based on the belief computed in the previous time interval, is also considered.

The final result of the proposed methodology is the posterior probability. In the implementation shown in this section, it consists of a local grid, where each cell represents a portion of the space and their values represent the probability of the cell is occupied by people. To make this information useful in a real-world application, it needs to be processed to decide if people were in fact detected and, in case they were detected, what are their locations. In this paper, we call this post-processing step as classification. Several classification methods could be used, including a simple fixed threshold, where it is assumed that there are people in a cell if its correspondent probability is larger than a given value. We could also have more sophisticated detections that, for example, would look for regions with given properties, such as area and velocity. In this context, we call these regions as blobs and blob detection refers to detecting regions that differ in properties compared to their neighborhood regions.

In this paper, for illustrative purposes, we propose a classification algorithm based on similar probabilities. The first step of this algorithm is the creation of blobs, which begins with the selection of cells whose probability value is greater than a threshold (the value 0.26 was used). Selected cells that are neighboring each other become part of the same blob. In the next step, all cells that do not belong to a blob, but that are neighbors of a blob, are visited. If the difference between the probability values of these cells and their neighboring cells in the blob is smaller than a threshold (in our case, 0.40), the cell is also included in the blob. Otherwise, a new blob is created. The procedure continues until all cells visited are part of a blob. A final step is required for selecting the blobs that represent people. The blobs selected are those that have an average probability value larger than the average of the blobs in the neighborhood, the probability of all cells are larger than 0.1 and have a maximum size of three cells or, if the blobs contain more than 3 cells, they fulfill all of the following criteria: (i) the blob may not contain more than 3 cells at the borders of the grid; (ii) the height and width of the blob is smaller than 8 cells and (iii) the ratio between the height and width of the blob is smaller than a threshold (0.5). To illustrate the classification step, we applied this algorithm in the semantic grid of Fig. 6r and show the result in Fig. 7. In this figure, people is represented by red cells. Movies that allows a qualitative evaluation of the proposed methodology followed by this classification procedure can be found in http://coro.cpdee.ufmg.br/movies/peopledetection.

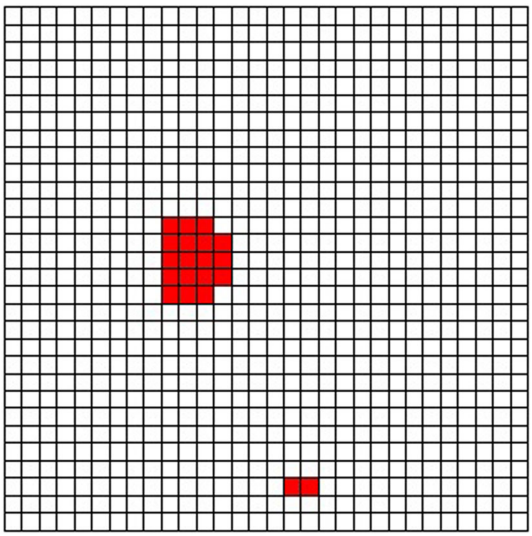


Fig. 7 Classification applied to the semantic grid of Fig. 6r

Table 2 Results of people detection. The detectors compared are: Detector 1 (Bellotto and Hu 2009), Detector 2 Spinello and Siegwart 2008, Detector 3 (Dollar et al. 2014) and the classification that follows the proposed approach

Detector	Recall	Precision	FPCPF
Detector 1	0.36	0.81	0.5
Detector 2	0.29	0.45	1.5
Detector 3	0.68	0.93	5.9
Proposed methodology with classification	0.72	0.86	3.3

To enable a quantitative evaluation of the proposed approach, all steps of the approach were executed followed by classification based on blob detection in a subset of the experiment's data containing 2044 iterations. This subset allows an assessment of different situations witnessed by the robot, which was moving with changes in its linear and angular velocities. Among this subset of images, images containing people with feet occluded were removed, since the mapping from the image frame to the grid explicitly considers the position of people's feet. Images with partial occlusions of other body parts were maintained. At the end, a total of 1094 iterations were selected and a ground truth constructed. The ground truth was obtained by manually annotating the image frames. This process found 1052 people detections on the 1094 images. Table 2 presents numerical results on this data set.

The results of the two laser-based detectors, the image-based detector, and the classification based on blob detection using the proposed methodology were automatically compared with the image ground truth projected on the grid. The following metrics were used to evaluate the experimental results: recall, which is the percentage of correct detections

with respect to the real number of persons; precision, defined before as the ratio between the number of correct detections and the total number of detections; and false-positive cells per frame (FPCPF) which measures the precision in location of the detections by computing the average number of cells that the detector considered occupied by people but that are free in the ground truth (the lower is the FPCPF, the better is the result).

The proposed approach combined with classification in comparison with all detectors showed an increase in recall, which means that the number of false negatives was decreased and a larger number of people actually in the scene were detected. The precision value remained among the two highest values of individual detectors. While Detector 3 presented a higher precision than the proposed methodology, its FPCPF was the worse among all detectors, indicating that it presents errors in the positioning of people, as expected. These results are in accordance with the idea that fusion makes the detection of persons more accurate than detection using single detectors. Notice that this is not a particular contribution of the proposed methodology, since it consists one of the main advantages of sensor integration. The results presented are compatible with the state-of-the-art of people detection with respect to precision and recall (Premebida et al. 2014; Wu et al. 2011; Araújo et al. 2011; Oliveira et al. 2010). It is important to notice that these results could be improved if better detectors are used. Another observation is that our methodology followed by classification gets higher recall and a small number of false positives per frame (FAPF = 0.12 for a recall of 0.72), demonstrating that it is possible to obtain a larger number of detections of the people who are on the scene keeping the number of false detections low.

The experimental results on this section indicate that, even with sensor motion, the proposed methodology may provide a reliable result that can be used as a basic step for several applications, such as tracking, navigation, people counting, human–robot interaction, and event detection. To show an example of application for our methodology, we have used the final occupancy grid computed by our method to track a person that is moving in the field of view of the robot's sensors. For this task, we have used a sequence of sensorial information that includes the snapshots in Fig. 9, where a person was tracked by several sample times, even when she is not detected by any of the sensors.

The tracked person starts her motion and appears into the grid as soon as she enters the FOV of the laser. She keeps moving forward in the corridor and eventually enters the FOV of the camera. Tracking began with the manual selection of the blob that represents the person on the grid. The next blobs were automatically selected in the grid. The criterion for selecting a blob is based on its position relative to the centroid of the blob selected at the previous instant of time. If there are more than one blob near the centroid, the one with the highest posterior probability is chosen. The centroid of the blob is selected as the cell whose coordinates are the average of the coordinates of the cells with the higher probability in the blob. If there is no blob close to the centroid of the previous blob, a blob that overlaps any of the cells of the previous blob is chosen or, as the last option, a blob in the 8-neighborhood of some cell of the previous blob is chosen.

Figure 8a shows the results of tracking for 111 consecutive iterations. In this figure, the centroid of the selected blobs is shown in black. For comparison, the cells where the person passed, according to the manually obtained laser ground truth, are shown in red in Fig. 8. One can notice by comparing Fig. 8a, b that the tracking on the grid was successful. During the tracking, the people detectors based on laser and images failed simultaneously to detect the person in six instants of time. Detector 1 found the person in 77 % of the iterations,

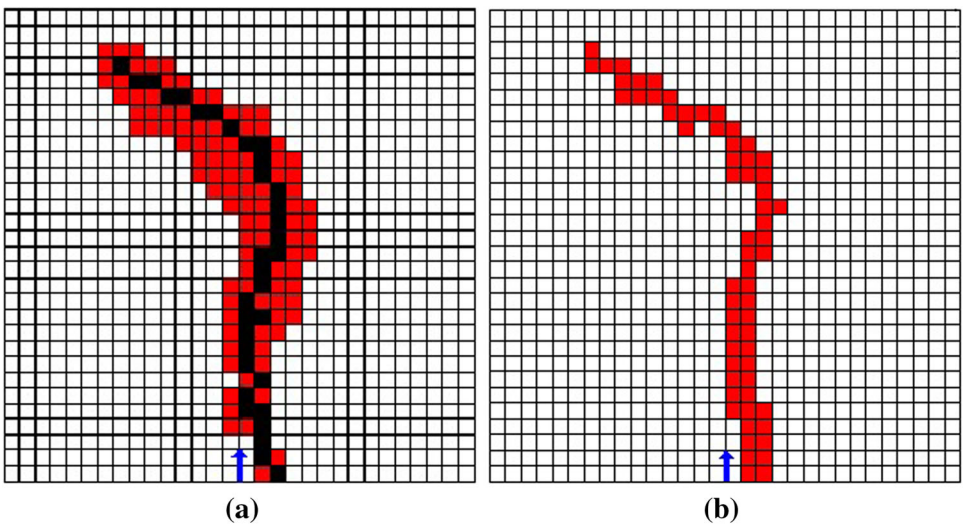


Fig. 8 Results of people tracking on the grid. The *arrow* indicates the beginning of the tracking. **a** *Black cells* represent the centroids of selected blobs (*red cells*). **b** Laser ground truth (Color figure online)

Fig. 9 Image sequence during the tracking experiment of Fig. 8a (a picture at every four iterations is shown). From *left to right*, *top to bottom* the tracked person crosses the grid starting in a position out of the field of view (FOV) of the camera. In the first frames she cannot be detected by the camera but she is already in the FOV of the laser

while detectors 2 and 3 detected the person 30 and 79 % of the time, respectively. This demonstrated the robustness of the approach in situations where the detectors fail (Fig. 9).

5 Conclusions

This paper proposed a Bayesian methodology for combining people detectors. To illustrate the approach, experiments with real data obtained using a laser rangefinder and a camera fixed on a mobile robot were executed in a dynamic environment. The fusion of three different people detectors indicated the improvements brought by the approach in relation to a single detector alone. In the iteration show in Fig. 6, for example, one can see that the final result shown in Fig. 6r is (i) more complete than the information obtained by each of the individual detectors, which fail simultaneously to detect one person on the scene, (ii) more precise in relation to detection, once it does not incorporate the false detection given by one of the detectors (Fig. 6o), and (iii) more accurate in relation to positioning, given that it refines the result obtained by the camera with laser information.

In general, the experimental results using a local semantic grid to represent the environment, which includes a people tracking application, show an improvement in relation to the state-of-the-art of people detection, demonstrating that it is possible to obtain a larger number of detections of people on the scene keeping a low number of false alarms.

The proposed approach allows the use of several combinations of people detectors, and it is completely flexible in relation to the sensors used, what makes their application possible in many contexts. To be adapted to other detectors and sensors, the only requirement is the development of a good detection model. Another flexibility is in relation to the people motion model, which was proven to be important for the final results. In the motion model used in this paper, people have uniform probability to move in any direction. If the direction of the movement is controlled, by the presence of a corridor, for example, this information could be used to determine a more precise model.

References

Adarve, J., Perrollaz, M., Makris, A., & Laugier, C (2012). Computing occupancy grids from multiple sensors using linear opinion pools. In: Proceedings IEEE International Conference Robotics and Automation (pp. 4074–4079).

Antunes, M., Barreto, J., Premebida, C., & Nunes, U. (2012). Can stereo vision replace a laser rangefinder? In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 5183–5190).

Araujo, A. R., Caminhas, D. D., & Pereira, G. A. S. (2015). An architecture for navigation of service robots in human-populated office-like environments. In: Proceedings of the IFAC Symposium on Robot Control (submitted).

Araújo, R. L., Lacerda, V., Hernandes, A., Mendonca, A., & Becker, M. (2011). Classificação de pedestres usando câmera e sensor lidar. In: Anais do Simpósio Brasileiro de Automação Inteligente (pp. 416–420).

Baig, Q., Perrollaz, M., & Laugier, C. (2014). A robust motion detection technique for dynamic environment monitoring: A framework for grid-based monitoring of the dynamic environment. *IEEE Robotics Automation Magazine*, *21*(1), 40–48.

Bellotto, N., & Hu, H. (2009). Multisensor-based human detection and tracking for mobile service robots. *IEEE Trans on Systems, Man, and Cybernetics*, *39*(1), 167–181.

Benenson, R., Omran, M., Hosang, J., & Schiele, B. (2015). Ten years of pedestrian detection, what have we learned? In: Computer Vision - ECCV 2014 Workshops, Lecture Notes in Computer Science, vol 8926 (pp. 613–627). Springer International Publishing.

Bota, S., & Nedesvchi, S. (2008). Multi-feature walking pedestrians detection for driving assistance systems. *Intelligent Transport Systems, IET*, *2*(2), 92–104.

Broggi, A., Cerri, P., Ghidoni, S., Grisleri, P., & Jung, H. G. (2009). A new approach to urban pedestrian detection for automatic braking. *IEEE Trans on Intelligent Transportation Systems*, *10*(4), 594–605.

Ceccarelli, M. (2011). Problems and issues for service robots in new applications. *International Journal of Social Robotics*, *3*(3), 299–312.

Cho, H., Seo, Y. W., Vijaya Kumar, B., & Rajkumar, R. (2014). A multi-sensor fusion system for moving object detection and tracking in urban driving environments. In: Proceedings of IEEE International Conference on Robotics and Automation (pp. 1836–1843).

Cui, J., Zha, H., Zhao, H., & Shibasaki, R. (2005). Tracking multiple people using laser and vision. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 2116–2121).

Daamen, W., & Hoogendoorn, S. (2007). Free speed distributions—based on empirical data in different traffic conditions. In: N. Waldau, P. Gattermann, H. Knoflacher, & M. Schreckenberg (Eds.), *Pedestrian and Evacuation Dynamics 2005* (pp. 13–25). Berlin: Springer.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol 1 (pp. 886–893).

Dollar, P., Appel, R., Belongie, S., & Perona, P. (2014). Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(8), 1532–1545.

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(4), 743–761.

Elfes, A. (1990). Occupancy grids: A stochastic spatial representation for active robot perception. In: Proceedings of Conference on Uncertainty in Artificial Intelligence (pp. 136–146). AUAI Press.

Geronimo, D., Lopez, A., Sappa, A., & Graf, T. (2010). Survey of pedestrian detection for advanced driver assistance systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(7), 1239–1258.

Gidel, S., Blanc, C., Chateau, T., Checchin, P., & Trassoudaine, L. (2009). Non-parametric laser and video data fusion: Application to pedestrian detection in urban environment. In: Proceedings of International Conference on Information Fusion (pp. 626–632).

Hofmann, M., Kaiser, M., Aliakbarpour, H., & Rigoll, G. (2011). Fusion of multi-modal sensors in a voxel occupancy grid for tracking and behaviour analysis. In: Proceedings of International Workshop on Image Analysis for Multimedia Interactive Services.

Hogenboom, M. (2013). Secret of Usain Bolt's speed unveiled. http://www.bbc.co.uk/news/science-environment-23462815. Accessed 06 Nov 2014.

Huerta, I., Ferrer, G., Herrero, F., Prati, A., & Sanfeliu, A. (2014). Multimodal feedback fusion of laser, image and temporal information. In: Proceedings of International Conference on Distributed Smart Cameras (pp. 25:1–25:6). ACM, New York, USA.

Kassir, A., & Peynot, T. (2010). Reliable automatic camera-laser calibration. In: Proceedings of the Australasian Conference on Robotics & Automation (p. 10).

Liu, Z., & von Wichert, G. (2014). Extracting semantic indoor maps from occupancy grids. *Robotics and Autonomous Systems*, *62*(5), 663–674.

MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., & Franklin, A. B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, *84*(8), 2200–2207.

Monteiro, G., Premebida, C., Peixoto, P., & Nunes, U. (2006). Tracking and classification of dynamic obstacles using laser range finder and vision. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems.

Ngako Pangop, L., Chausse, F., Chapuis, R., & Cornou, S. (2008). Asynchronous Bayesian algorithm for object classification: Application to pedestrian detection in urban areas. In: Proceedings of International Conference on Information Fusion (pp. 1–7).

Nüchter, A., & Hertzberg, J. (2008). Towards semantic maps for mobile robots. *Journal of Robotics and Autonomous Systems*, *56*(11), 915–926.

Oliveira, L., Nunes, U., Peixoto, P., Silva, M., & Moita, F. (2010). Semantic fusion of laser and vision in pedestrian detection. *Pattern Recognition*, *43*, 3648–3659.

Papoulis, A., & Pillai, S. U. (2002). *Probability, random variables, and stochastic processes* (4th ed.). New York: Mc-Graw Hill.

Pereira, F. G., Vassallo, R. F., & Salles, E. O. T. (2013). Human-robot interaction and cooperation through people detection and gesture recognition. *Journal of Control, Automation and Electrical Systems*, *24*(3), 187–198.

Premebida, C., Carreira, J., Batista, J., & Nunes, U. (2014). Pedestrian detection combining RGB and dense LIDAR data. In: Proceedings of International Conference on Intelligent Robots and Systems (pp. 4112–4117).

Premebida, C., Ludwig, O., & Nunes, U. (2009). Lidar and vision-based pedestrian detection system. *Journal of Field Robotics*, *26*(9), 696–711.

SICK. (2006). Technical description for the lms200/211/221/291 laser measurement systems. Tech. rep., SICK AG Waldkirch, Germany.

Spinello, L., & Siegwart, R. (2008). Human detection using multimodal and multidimensional features. In: Proceedings of the IEEE International Conference on Robotics and Automation (pp. 3264–3269).

Stein, G., Mano, O., & Shashua, A. (2003). Vision-based acc with a single camera: bounds on range and range rate accuracy. In: Proceedings of IEEE Intelligent Vehicles Symposium (pp. 120–125).

Thrun, S., Burgard, W., & Fox, D. (2005). *Probabilistic robotics*. Cambridge, MA: The MIT Press.

Utasi, A., & Benedek, C. (2013). A Bayesian approach on people localization in multicamera systems. *IEEE Transactions on Circuits and Systems for Video Technology*, *23*(1), 105–115.

Varga, R., Vesa, A., Jeong, P., & Nedevschi, S. (2014). Real-time pedestrian detection in urban scenarios. In: Proceedings of IEEE International Conference on Intelligent Computer Communication and Processing (pp. 113–118).

Varvadoukas, T., Giotis, I., & Konstantopoulos, S. (2012). Detecting human patterns in laser range data. In: Proceedings of the European Conference on Artificial Intelligence, vol 242 (pp. 804–809).

Wu, B., Liang, J., Ye, Q., Han, Z., & Jiao, J. (2011). Fast pedestrian detection with laser and image data fusion. In: Proceedings of the International Conference on Image and Graphics (pp. 605–608).

Yguel, M., Aycard, O., & Laugier, C. (2006). Efficient gpu-based construction of occupancy girds using several laser range-finders. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 105–110).

Yoder, J.D., Perrollaz, M., Paromtchik, I., Mao, Y., & Laugier, C. (2010). Experiments in vision-laser fusion using the bayesian occupancy filter. In: Proceedings of International Symposium on Experimental Robotics, Delhi, India.