

# Chapter 1

## People Detection using Multiple Sensors on a Mobile Robot

Zoran Zivkovic and Ben Kröse

### 1.1 Introduction

<sup>1</sup> Robots are moving out of laboratories into public places where the human beings have to be taken into account. Such robots should be able to interact with humans and show aspects of human style social intelligence. This also implies that in addition to the perception required for the conventional functions (localization, navigation, etc.), a "socially interactive" robot needs strong human oriented perceptual capabilities [1, 16]. For a start the robot should be able to accurately and robustly detect and localize the persons around it.

Person detection from images is a widely studied problem in the computer vision research area. Two types of applications can be distinguished. The first type is surveillance where usually much knowledge is available about the environment, camera position and camera parameters. This knowledge provides additional cues for person detection. For example in most man made environments people walk over a floor plane which leads to a limited set of possible person position in an image. Furthermore, the camera is often static and this can help to distinguish persons from the static background. The second type of application considers a more general and difficult problem where not much a priori knowledge is available about the images, e.g. images or videos from the internet. A common approach in such situations is to use the whole image to infer more about the environment and the camera which can then help to detect people, e.g. [10].

Typical robotics applications differ from the typical computer vision applications in a number of aspects. First, robotics systems are usually equipped with multiple sensors. For example 2D laser range scanner is often used to detect persons legs.

---

Zoran Zivkovic and Ben Kröse  
University of Amsterdam, Intelligent Systems Laboratory, Kruislaan 403, 1098SJ Amsterdam, The Netherlands e-mail: {zivkovic,krose}@science.uva.nl

<sup>1</sup> In D.Kragic,V.Kyrki,eds: Unifying Perspectives in Computational and Robot Vision, Springer, 2008.

Properly combining the information from different sensors can improve the detection results. Second, similar to the surveillance applications, camera and other sensors positions and parameters are usually known. However, the sensors are not static since they are mounted on a moving platform, a mobile robot or a vehicle. Finally, the fact that robots move can be an advantage. Actively moving the sensors can improve the detection results, for example moving closer to the object or viewing it from another view point.

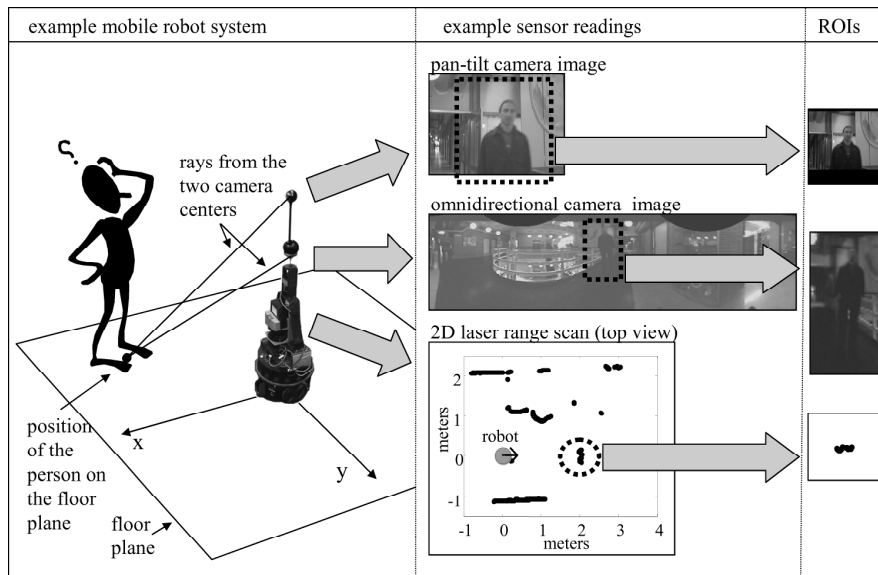
This chapter considers the important problem of dealing with multiple sensors. An approach for combining information from multiple sensors for people detection on a mobile robot is described. A person will be represented by a constellation of body parts. Person body parts are detected and the parts are constrained to be at certain positions with respect to each other. Similar part based representations are widely used in the computer vision area for describing objects in images. A probabilistic model is presented here to combine part detections from multiple sensors typical for mobile robots. For detecting the body parts specific detectors can be constructed in many ways. In this chapter the Ada-Boost [7] is used as a general "out of box" approach for building the part detectors.

The chapter starts with the related work which is presented in Section 1.2. Next, in Section 1.3 people detection using 2D laser range scanner is considered. Persons legs can be detected in the scans. A probabilistic part-based representation is presented that takes into account the spatial arrangement of the detected legs. The method is inspired by the latest results on the "part-based representations" from the computer vision area and the work of Weber, Perona and colleagues [21, 5]. The approach takes into account that the leg detector might produce false detections or fail to detect legs, for example because of partial occlusion. Section 1.4 describes a straightforward way to extend the presented probabilistic model to properly combine body parts detected using other sensors that might be present on the robot, a pan-tilt camera and an omnidirectional camera in our case, see Figure 1.1. Evaluation of the proposed model and some practical issues are discussed in Section 1.5. Finally, the conclusions are given in Section 1.6.

## 1.2 Related work

A 2D laser range scanner is often used in robotics for detecting and tracking people [13, 11]. People are detected by finding their legs in the laser scans. Disadvantages of using the laser scans for people detection are: the persons can be detected only at limited distances from the robot, low detection rate in highly cluttered environments and that the methods fail when the person legs are occluded. Other sensors were also used like thermal vision [17], stereo vision [9] and regular cameras [23].

Person detection from images is a widely studied problem in the computer vision area. Many of the presented algorithms aim at the surveillance applications [6] and are not applicable to mobile platforms since they assume static camera. There is also



**Fig. 1.1** Example moving robot platform equipped with three sensors: a 2D laser range scanner, a pan-tilt camera and an omnidirectional camera. The sensors are calibrated and their pose with respect to the floor plane is known. Given the typical size of a person we can define a region of interest (ROI) in each sensor corresponding to a floor plane position as shown.

a large number of papers considering the people detection without the static camera assumption, e.g. [8, 14, 22, 15].

The people detection can be seen as a part of the more general problem of object detection. Many approaches were considered in the computer vision area. Recently it was shown that fast and reliable detection can be archived using "out of box" technique Ada-Boost to build classifiers [7]. For example Haar-like features with Ada-Boost were successively used for face detection by Viola and Jones [19]. Similar techniques were used for people detection [20, 18].

Another approach for object detection in images is the so called "part-based representation". Various part-based representations, e.g. [2, 5, 3, 4], are demonstrated to lead to high recognition rates. An important advantage of the part-based approach is it relies on object parts and therefore it is much more robust to partial occlusions than the standard approach considering the whole object.

The part-based people detection was considered a number of times. Seemann et al. [14] use SIFT based part detectors but do not model part occlusions. Wu and Nevatia [22] describe the part occlusions but the occlusion probabilities and part positions are learned in a supervised manner. We base our algorithm on a principled probabilistic model of the spatial arrangement of the parts similar to the work of Weber, Perona and colleagues [21, 5]. An advantage of having a proper probabilistic model is that, after constructing the part detectors, the part arrangement and occlusion probabilities can be automatically learned from unlabelled images.

This chapter presents a part-based approach and shows how it can be used to properly combine information from multiple sensors on a mobile robot, 2D range data, omnidirectional camera and pan-tilt camera in our case.

### 1.3 Part based model

Legs of a person standing in front of a robot can be detected using a 2D laser range scanner [11]. A part-based model is presented that takes into account the possible distance between the detected persons legs. The fact that leg detector might produce false detections or fail to detect legs, for example because of partial occlusion, is taken into account. The model also presents the base for combining information from different sensors as described later.

#### 1.3.1 Part detection

A human is detected by detecting  $P$  human body parts, in this case  $P = 2$  for the legs. The 2D position of a leg is  $\mathbf{x}_p = (x_p, y_p)$ . The Gaussian distribution is used as a simple model of the leg positions:

$$p_{shape}(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (1.1)$$

where  $\mathbf{x} = (\mathbf{x}_1 \dots \mathbf{x}_P)$  is a  $2P$  long vector containing all the 2D part positions,  $\boldsymbol{\mu}$  is the mean and  $\boldsymbol{\Sigma}$  is a  $(2P) \times (2P)$  covariance matrix. If the covariance matrix is diagonal than this model can be seen as describing "string-like" constraints between the body-part positions [4]. The non-diagonal covariance matrix will express additional relations between the positions of the body parts.

A laser range scan is first divided into segments by detecting abrupt changes using the Canny edge detector. Reliable leg detection is performed using a set of geometric features and Ada-Boost classifier as described in [24]. Let  $N$  denote the number of segments classified as legs and let  $\mathbf{x}_j$  denote the 2D position of the  $j$ -th detection. All leg detections from one scan are given by:

$$\mathcal{X} = (\mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_N) \quad (1.2)$$

The 2D image position  $\mathbf{x}_j = (x_j, y_j)$  of the  $j$ -th detection is calculated as the mean position of the 2D scan segment points. Note that sometimes the legs cannot be separated or their appearance in the scan might change drastically. Furthermore, in cluttered environments many other objects, e.g. chairs or tables, may produce 2D scan output similar to human legs and some detections might be false detections.

### 1.3.2 Missing detections and clutter

From a range scan the collection of person's leg candidates  $\mathcal{X}$  is extracted but some of them are true and some false detections. To indicate which detections are correct a  $P = 2$  element vector  $\mathbf{h}$  is used with element  $h_p = j$ ,  $j > 0$ , indicating that the  $j$ -th detection  $\mathbf{x}_j$  belongs to the of the  $p$ -th body part (leg) and the other detections of that part are false detections. Given  $\mathbf{h}$  the 2D positions of the person's legs are composed of the corresponding detections  $\mathbf{x} = (\mathbf{x}_{h_1} \mathbf{x}_{h_2})$ . The set of all other detections that belong to the background clutter are denoted by  $\mathbf{x}^{bg}$ .

It is possible that a leg was not detected indicated using  $h_p = 0$ . The position of a not detected leg is considered as missing data. To make distinction between the missing and the observed parts the set of missing parts is denoted as  $\mathbf{x}^m$  and the set of observed parts as  $\mathbf{x}^o$ . To indicate the fact that there can be missing parts, the probabilistic model of the arrangement of the body parts (1.1) will be written as:  $p_{shape}(\mathbf{x}) = p_{shape}(\mathbf{x}^o, \mathbf{x}^m)$ .

### 1.3.3 Probabilistic model

The the possibility of part detector false alarms and missed detections of body parts of a person is determined by the unknown assignment hypotheses vector  $\mathbf{h}$ . The probabilistic model can be written as a joint distribution:

$$p(\mathcal{X}, \mathbf{x}^m, \mathbf{h}) = p(\mathcal{X}, \mathbf{x}^m | \mathbf{h}) p(\mathbf{h}) \quad (1.3)$$

where both  $\mathbf{x}^m$  and  $\mathbf{h}$  are unknown missing data.

Two auxiliary variables  $\mathbf{b}$  and  $\mathbf{n}$  are used to further define  $p(\mathbf{h})$ . The variable  $\mathbf{b} = \text{sign}(\mathbf{h})$  is a binary vector that denotes which parts have been detected and which not. The value of the element  $n_p \leq N_p$  of the vector  $\mathbf{n}$  represents the number of detections of part  $p$  that are assigned to the background clutter. The joint distribution (1.3) becomes:

$$p(\mathcal{X}, \mathbf{x}^m, \mathbf{h}, \mathbf{n}, \mathbf{b}) = p(\mathcal{X}, \mathbf{x}^m | \mathbf{h}) p(\mathbf{h} | \mathbf{n}, \mathbf{b}) p(\mathbf{n}) p(\mathbf{b}) \quad (1.4)$$

where  $\mathbf{b}$  and  $\mathbf{n}$  are assumed to be independent and:

$$p(\mathcal{X}, \mathbf{x}^m | \mathbf{h}) = p_{shape}(\mathbf{x}^o, \mathbf{x}^m) p_{bg}(\mathbf{x}^{bg}) \quad (1.5)$$

where the observed parts  $\mathbf{x}^o$ , the missing parts  $\mathbf{x}^m$  and the false detections from clutter  $\mathbf{x}^{bg}$  correspond to the hypothesis  $\mathbf{h}$ . The  $p_{bg}(\mathbf{x}^{bg})$  is the distribution of the false detections usually uniform or a wide Gaussian.

The probability  $p(\mathbf{b})$  describing the presence or absence of parts is modelled as an explicit table of joint probabilities. Each part can be either detected or not, so there are in total  $2^P$  possible combinations that are considered in  $p(\mathbf{b})$ .

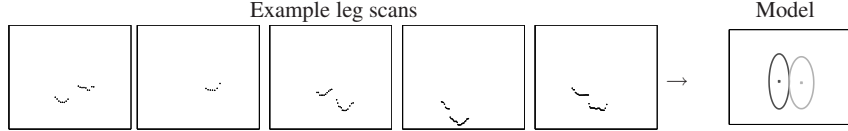
The background part detections are assumed independent of each other and the number of detections  $\mathbf{n}$  is modelled using Poisson distribution with mean  $M_p$  [21]. Different  $M_p$ -s for different parts admit different detector statistics. The Poisson parameter will be denoted by vector  $\mathbf{M} = (M_1 \dots M_P)$ .

The density  $p(\mathbf{h}|\mathbf{n}, \mathbf{b})$  is defined as:

$$p(\mathbf{h}|\mathbf{n}, \mathbf{b}) = \begin{cases} 1/|\mathcal{H}(\mathbf{b}, \mathbf{n})| & \text{if } \mathbf{h} \in \mathcal{H}(\mathbf{b}, \mathbf{n}), \\ 0 & \text{otherwise.} \end{cases} \quad (1.6)$$

where  $\mathcal{H}(\mathbf{b}, \mathbf{n})$  is the set of all hypotheses consistent with the values of  $\mathbf{b}$  and  $\mathbf{n}$ . Here  $|\mathcal{H}(\mathbf{b}, \mathbf{n})|$  denotes the total number all consistent part assignment hypotheses. This expresses that these hypotheses are considered equally likely.

### 1.3.4 Learning model parameters



**Fig. 1.2** Example person's legs scans from the data set used to train the probabilistic part-based model and the learned model parameters. For each part its mean position contained in the parameter  $\mu$  is presented. The ellipse represents the 1-sigma uncertainty of the part position as described by the diagonal elements of the covariance matrix  $\Sigma$ .

The density distribution (1.4) will have the following set of parameters  $\Omega = \{\mu, \Sigma, p(\mathbf{b}), \mathbf{M}\}$ :

$$p(\mathcal{X}, \mathbf{x}^m, \mathbf{h}) = p(\mathcal{X}, \mathbf{x}^m, \mathbf{h}|\Omega) \quad (1.7)$$

The likelihood of a collection of detected parts  $\mathcal{X}$  is obtained by integrating over the hidden hypotheses  $\mathbf{h}$  and the missing parts:

$$p(\mathcal{X}|\Omega) = \sum_{\text{all possible } \mathbf{h}} \int_{\mathbf{x}^m} p(\mathcal{X}, \mathbf{x}^m, \mathbf{h}|\Omega). \quad (1.8)$$

Integrating over the missing parts  $\mathbf{x}^m$  for the Gaussian distribution can be performed in closed form.

To estimate the parameters of the model a set of  $L$  aligned scans of persons is used. The collection of leg detections for  $i$ -th scan will be denoted as  $\mathcal{X}_i$ . The maximum likelihood estimate of the parameters  $\Omega$  is computed by maximizing the likelihood of the data:

$$\prod_i^L p(\mathcal{X}_i|\Omega) \quad (1.9)$$

using expectation maximization algorithm, see [21] for details.

### 1.3.5 Detection

Let us denote the maximum likelihood parameters learned from a set of scans of persons as  $\Omega_{person}$ . For a set of scans from the office clutter the  $p_{bg}(\mathbf{x}^{bg})$  and other parameters can be estimated, denoted as  $\Omega_{bg}$ . Given a new scan and extracted the set of detected parts  $\mathcal{X}$ . The scan is either a scan of a person or some background clutter:

$$p(\mathcal{X}) = p(\mathcal{X}|Person)p(Person) + p(\mathcal{X}|BG)p(BG) \quad (1.10)$$

where  $p(Person)$  and  $p(BG)$  are unknown a priori probabilities that the scan contains a person or background. The a posteriori probability that there is a person is:

$$p(Person|\mathcal{X}) = \frac{p(\mathcal{X}|Person)p(Person)}{p(\mathcal{X})} \approx \frac{p(\mathcal{X}|\Omega_{person})p(Person)}{p(\mathcal{X}|\Omega_{person})p(Person) + p(\mathcal{X}|\Omega_{bg})p(BG)} \quad (1.11)$$

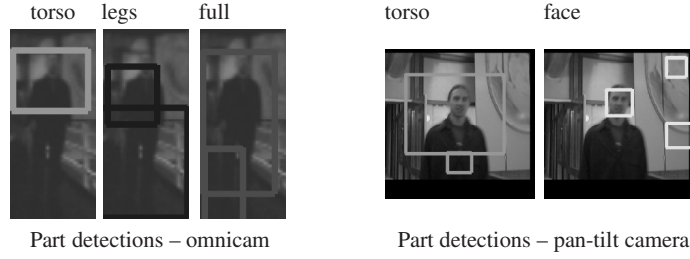
The last step above is an approximation since the maximum likelihood estimates for the model parameters  $\Omega_{person}$  and  $\Omega_{bg}$  are used instead of integrating over all possible parameter values. Calculating  $p(\mathcal{X}|\Omega)$  is done using (1.8).

## 1.4 Combining multiple sensors

Robots are often equipped with multiple sensors. For example an omnidirectional and a pan-tilt camera as in Figure 1.1. In this section the part based model from the previous section is extended to include part detections from the corresponding images.

### 1.4.1 Part detection in images

Haar-like-feature classifiers are used to detect various human body parts in images. Each classifier is trained using Ada-Boost algorithm on a large set of example images of the corresponding body part [19]. Here the classifiers are trained on face, upper body, lower body and full body images. The part detectors can lead to many false alarms and missed detections [12], see Figure 1.3.



**Fig. 1.3** Example body part detections with some false detections.

### 1.4.2 Extending the part-based model

The part based model from the previous section that was applied to the 2D range leg detections can be easily extended with the human body parts detected in the images. Instead of 2 parts there will be  $P = 2 + 3 + 2 = 7$  body parts and  $\mathbf{x} = (\mathbf{x}_1 \dots \mathbf{x}_P)$  is a  $2P$  long vector containing 2D leg positions, the 2D image positions for the upper body, lower body and full body detected in omniscam images and face and upper body detected positions from the pan-tilt camera.

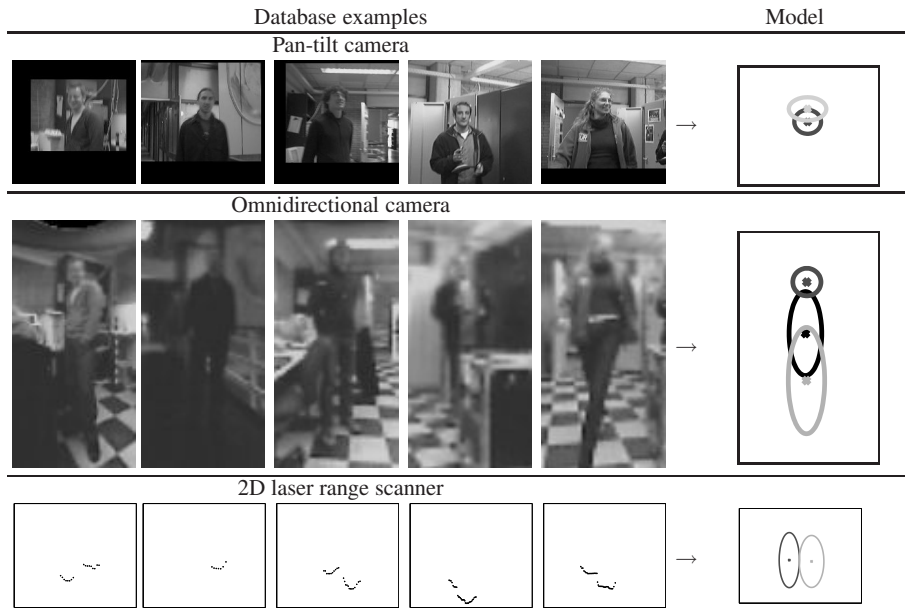
The positions of all detected parts are summarized in a data structure:

$$\mathcal{X} = \begin{pmatrix} \mathbf{x}_{1,1} & \mathbf{x}_{1,2} & \dots & \mathbf{x}_{1,N_{leg1}} & & & \\ \mathbf{x}_{2,1} & \mathbf{x}_{2,2} & \dots & \mathbf{x}_{2,N_{leg2}} & & & \\ \mathbf{x}_{3,1} & \mathbf{x}_{3,2} & \dots & \dots & \mathbf{x}_{3,N_{up-body}} & & \\ \mathbf{x}_{4,1} & \mathbf{x}_{4,2} & \dots & \mathbf{x}_{4,N_{low-body}} & & & \\ \mathbf{x}_{5,1} & \mathbf{x}_{5,2} & \dots & \dots & \mathbf{x}_{5,N_{full-body}} & & \\ \mathbf{x}_{6,1} & \mathbf{x}_{6,2} & \dots & \dots & \mathbf{x}_{6,N_{face-pan-tilt}} & & \\ \mathbf{x}_{7,1} & \mathbf{x}_{7,2} & \dots & \mathbf{x}_{7,N_{up-body-pan-tilt}} & & & \end{pmatrix} \quad (1.12)$$

with one row per part and where each row contains information about the detections of the corresponding body part. The first two rows are repeated since the same detector is used for both legs detected in the range scans. The element  $\mathbf{x}_{p,j}$  contains the 2D positions for the legs or the 2D image position for the parts detected in images of the  $j$ -th detection of the  $p$ -th part. The rows of  $\mathcal{X}$  can have different lengths and some might be empty if that part was not detected.

Again the hidden  $P$  dimensional assignment vector  $\mathbf{h}$  is used with element  $h_p = j$ , indicating that the  $j$ -th detection of the  $p$ -th part  $\mathbf{x}_{p,j}$  belongs to the object and other detections of that part are false detections. Given  $\mathbf{h}$  the shape of the object is composed of the corresponding detections  $\mathbf{x} = (\mathbf{x}_{1,h_1} \dots \mathbf{x}_{P,h_P})$ . Note that since the same detector is used for both legs, care should be taken not to select the same leg detection for both legs.

The other model equations remain the same and the same procedure can be used to learn now the part based model containing the part detectors from both sensors. Example part based model learned from multiple sensors is presented in Figure 1.4.



**Fig. 1.4** Examples from the data set used to train the probabilistic part-based model and examples of learned part arrangement model parameters. For each part its mean position contained in the parameter  $\mu$  is presented. The ellipse represents the 1-sigma uncertainty of the part position as described by the diagonal elements of the covariance matrix  $\Sigma$ .

## 1.5 Experiments

The presented method for combining information from multiple sensors is evaluated here. Building reliable part detectors for each sensor is considered first.

### 1.5.1 Part detection

A data set of 2D range scans was recorded to build the leg detector. The URG-04LX 2D range scanner was mounted on our robot at 50cm above the floor. A set of 3530 scans was recorded while driving the robot through the corridors and cluttered offices in our building. This gives in total 4032 scan segments corresponding to person's legs and 14049 segments from the background clutter.

For each scan segment we extract the set of 12 geometric features, such as segment size, curvature, see [24] for details. The Gentle AdaBoost algorithm [7] then automatically selects the relevant features during training. Separate features are used to build simple linear classifiers, the so called "weak" classifiers. The classifiers are then combined to form the final classifier. The final classifier contains 25 classifiers.

The number of classifiers was chosen to such that recognition results do not improve for adding more classifiers. Note that some features might be selected a number of times. The final classifier leads to a reliable leg detection. However, leg detection in cluttered office environments remains difficult since many object can produce range scan similar to legs. Details are given in [24].

The Haar-like-feature based image part detectors we used in our experiments were trained on the MIT pedestrian data set [12] and are available in the Intel OpenCV library.

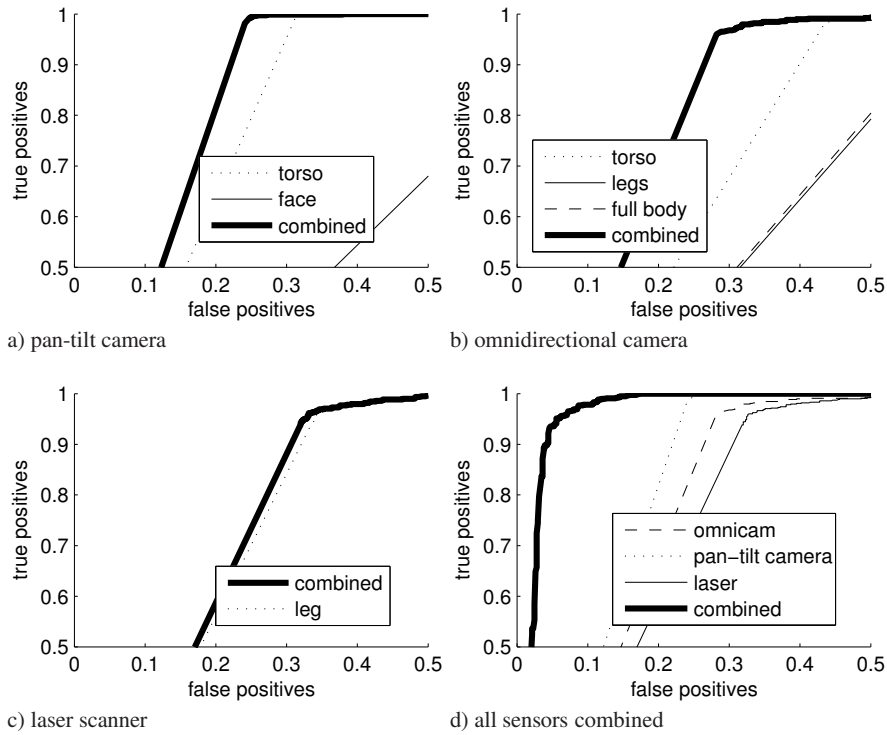
### 1.5.2 Multiple sensor people detection

For evaluating the part-arrangement model, we collected a realistic data set simulating the scenario where a human is introducing the robot to a new environment. Five persons were asked to lead the robot around our office environment. The robot was teleoperated. During the teleoperation the movements of the robot were such as to try to keep the person in the field of view of the pan-tilt camera. The data set contains 3200 images from the both cameras captured at 4 frames/second and the corresponding laser scans. On average each person was leading the robot for 2 – 3 minutes and there were around 600 images recorded for each person.

The calibrated omnidirectional camera images were used to manually select the ground truth person position on the floor-plane. The selected person position was used to cut out the corresponding regions from all three calibrated sensors, see Figure 1.1. The aligned images cut out of the omniscam images were  $56 \times 112$  pixels and the corresponding images from the pan-tilt camera were  $112 \times 112$  pixels, see Figure 1.4.

From the whole data set, 1000 randomly chosen images and scans were used to train our part based model and remaining part of the data set was used for testing. The automatically learned part based model parameters are presented in Figure 1.4. It can be observed that there is more uncertainty in the positions of the full and lower body than for the upper body region in the omniscam images. The pan-tilt camera was not very stable and it was shaking during the robot movements. This explains the larger uncertainty in the horizontal position of the detected face and upper body in the images from the pan-tilt camera.

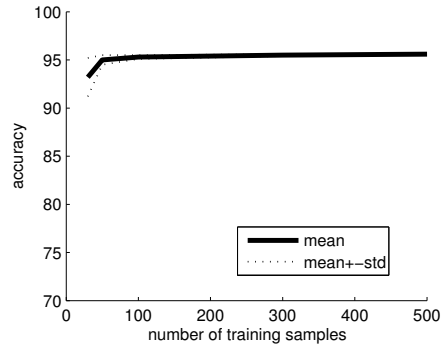
In order to test recognition results, a set of 5000 no-person parts of the images and scans are selected from the data set. The recognition results on this data set containing aligned images of people and corresponding scan segments are summarized in Figure 1.5. The results are presented as recognition Receiver Operating Characteristic (ROC) curves. Changing the a priori chances  $p(Person)$  and  $p(BG)$  in (1.11), various values are obtained for true positive (TP) and false positive (FP) detections used to plot the ROC curves. The ROC curves of the single part detectors for each sensor are also reported. Face detection in images is usually very reliable. However, for the pan-tilt camera images from our data set the face detector performs poorly, Figure 1.5a. This is mainly because the persons were not facing the robot very often



**Fig. 1.5** Recognition Receiver Operating Characteristic (ROC) curves

during the trials where they were leading the robot around our office environment. The persons were facing the robot at the start of each trial and also later on from time to time when they turned around to check if the robot is following them. Furthermore, it can be observed from the ROC curves that the combination of parts leads to much better results. The improvement is small for the laser scanner, Figure 1.5c. The largest improvement can be noted when the different sensor are combined, Figure 1.5d.

Learning the part detectors using the Ada-Boost requires often long time and many training examples [15]. On the other hand, once the part detectors are available, learning the part arrangement model usually does not require many training examples [5]. Learning the part arrangement model parameters for the 7 parts takes around 2 minutes for 1000 images in our Matlab implementation. In Figure 1.6 the recognition accuracy for the various sizes of the data set used to training the part arrangement model is presented. For each size of the training data set the experiments are repeated by choosing the training data randomly 10 times. The mean and the standard deviation of the maximum accuracy is presented. It can be observed that consistent high accuracy recognition can be achieved even with only 50 training data samples. In practice this means that given reliable part detectors for each



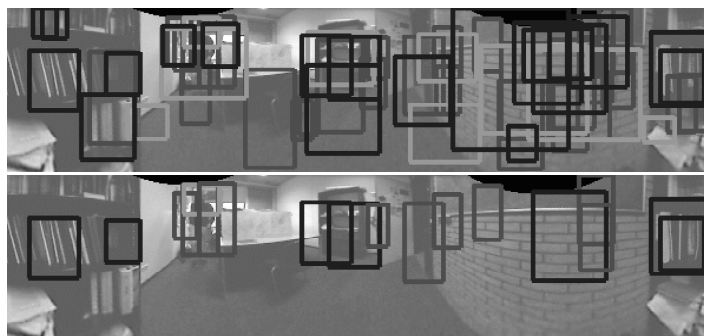
**Fig. 1.6** Recognition accuracy for various sizes of the dataset used to training the part arrangement model. Mean results and the standard deviation from 10 random trials is presented.

sensor, only a small additional effort needs to be made to construct the part-based combination of the detection results from multiple sensors.

### 1.5.3 Recognition from the robot

The part based model detection is implemented on our robot, see Figure 1.1. The assumption is made that the people walk over a flat ground floor surface - true in most man-made environments. A set of possible 2D floor positions  $\mathcal{T}_i$  is defined. In the experiments a  $10m \times 10m$  area around the robot is used and a grid of possible positions at every  $10cm$ . This gives 10000 possible floor points  $\mathcal{T}_i$  to evaluate the part based model. The sensors are calibrated and their pose with respect to the floor plane is known. Given the typical size of a person we can define a region of interest (ROI) in each sensor corresponding a floor plane position, see Figure 1.1. The data from the National Center for Health Statistics ([www.cdc.gov/nchs/](http://www.cdc.gov/nchs/)) is used. For adult humans, the mean height is 1.7m with a standard deviation of 0.085m. The maximal height of a human is taken to be the mean plus three standard deviations and the width to be  $1/2$  of the height. For each floor position  $\mathcal{T}_i$  we also extract the corresponding segments from the images and the range scan and use (1.11) to decide if there is a person at that floor position. Since (1.11) is computed at a dense grid of ground points, it often has large values for a number of ground points around the position where the person actually is. Therefore the persons are detected as the local maxima of (1.11).

The first stage of the algorithm where the body parts are detected in the omnidirectional images is the most computationally expensive. Running the three Haar-like-feature based part detectors on a  $600 \times 150$  panoramic image takes on average 400ms on a 2GHz PC. This is the time needed for checking every image position and all possible part sizes. The possible part sizes start from the initial part size and then the part size is increased 1.1 times until it gets out of the image borders. The floor



**Fig. 1.7** Body part detection in omniscam images (top) and the heavily reduced set of detections when the floor plane constraint is used (below).

constraint can heavily reduce the number of positions and part sizes to search and detection can be done in around 100ms, see Figure 1.7. Once the parts are detected, detecting persons using our model takes around 25ms. Currently, the people detection with all three sensors and 7 detected parts can be performed 5 times/second in our implementation on a 2GHz single processor.



Two correct detections of partially occluded people.



Two correct detections. The persons are in the dark and hardly visible



One correct and one false detection.

**Fig. 1.8** Example people detection results in panoramic images recorded from a moving robot.

In Figure 1.8 a few panoramic images with the detection results are presented to illustrate the typical detection results. The data set from the human following trials

was used to evaluate the actual detection performance on a mobile robot. A small subset of 100 annotated images and scans is used to train the model. The model is then applied using the floor constraint to detect people in the images and the range scans. The ground truth positions manually selected from the omniscam images were used to evaluate the performance. If a person was detected the corresponding rectangle ROI in the omniscam image was calculated, see Figure 1.8, and compared to the manually selected one using a relative overlap measure. Let  $R_{gt}$  be the image region defined by the ground truth bounding box. Let  $R_e$  be the estimated rectangle ROI (corresponding to the local maximum of (1.11)). The relative overlap is defined by:

$$overlap = \frac{R_e \cap R_{gt}}{R_e \cup R_{gt}} \quad (1.13)$$

where  $R_e \cap R_{gt}$  is the intersection and  $R_e \cup R_{gt}$  is the union of the two image regions. The relative overlap can have values between 0 and 1. A detection is considered to be true detection if the overlap was larger than 0.5. For the people following data set of 3200 sensor readings, there were 96% correctly detected people and only 120 false detections.

## 1.6 Conclusions

Due to the large variability in shape and appearance of different people the problem of people detection in images remains difficult even after many years of research [15]. The detection results can be improved if multiple sensors are combined. This chapter presents a people detection approach that combines information from a set of calibrated sensors. Mobile robots that often have various sensors are a typical example of a multisensory system. The approach is inspired by the part-based object representation from the computer vision area. A person is represented by a constellation of body parts. The person body parts are detected and the parts are constrained to be at certain positions with respect to each other. The presented probabilistic model combines the part detections from multiple sensors and can achieve person detection robust to partial occlusions, part detector false alarms and missed detections of body parts. The method is evaluated using a mobile test platform equipped with a pan-tilt camera, an omnidirectional camera and a 2D laser range scanner. The evaluation results show that highly reliable people detection can be achieved by properly combining the three sensors.

**Acknowledgements** This work has been sponsored by EU FP6-002020 COGNIRON ("The Cognitive Companion") project.

## References

1. Breazeal, C.: Designing sociable robots. MIT Press, Cambridge (2002)
2. Burl, M., Leung, T., Perona, P.: Recognition of planar object classes. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (1996)
3. Fei-Fei, L., Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (2005)
4. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. *Intl. Journal of Computer Vision* **61**(1), 55–79 (2005)
5. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (2003)
6. Ferryman, J., Crowley, J. (eds.): Proc. of the 9th IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (2006)
7. Friedman, J.H., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Technical Report, Dept. of Statistics, Stanford University (1998)
8. Gavrila, D., Philomin, V.: Real-time object detection for smart vehicles. In Proc. of the Intl. Conf. on Computer Vision (1999)
9. Giebel, J., Gavrila, D., Schnrr, C.: A Bayesian framework for multi-cue 3D object tracking. In Proc. of the European Conf. on Computer Vision (2004)
10. Hoiem, D., Efros, A., Hebert, M.: Putting objects in perspective. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (2006)
11. Arras, K., Mozos, O., Burgard, W.: Using boosted features for detection of people in 2D range scans. In Proc. of the IEEE Intl. Conf. on Robotics and Automation (2007)
12. Kruppa, H., Castrillon-Santana, M., Schiele, B.: Fast and robust face finding via local context. In: Proc of the IEEE Intl. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (2003)
13. Schulz, D., Burgard, W., Fox, D., Cremers, A.: People tracking with a mobile robot using sample-based joint probabilistic data association filters. *International Journal of Robotics Research* **22**(2), 99–116 (2003)
14. Seemann, E., Leibe, B., Mikolajczyk, K., Schiele, B.: An evaluation of local shape-based features for pedestrian detection. In Proc. of the British Machine Vision Conference (2005)
15. Munder, S., Gavrila, D.M.: An experimental study on pedestrian classification. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **28**(11), 1863–1868 (2006)
16. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. *Robots and Autonomous Systems* **42**, 143–166 (2003)
17. Treptow, A., Cielniak, G., Duckett, T.: Real-time people tracking for mobile robots using thermal vision. *Robotics and Autonomous Systems* **54**(9), 729–739 (2006)
18. Tuzel, O., Porikli, F., Meer, P.: Human detection via classification on riemannian manifolds. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (2007)
19. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (2001)
20. Viola, P., Jones, M., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In Proc. of the Intl. Conf. on Computer Vision (2003)
21. Weber, M., Welling, M., Perona, P.: Unsupervised learning of models for recognition. In Proc. of the European Conf. on Computer Vision (2000)
22. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In Proc. of the Intl. Conf. on Computer Vision (2005)
23. Zajdel, W., Zivkovic, Z., Kröse, B.: Keeping track of humans: have I seen this person before? In Proc. of the IEEE Intl. Conf. on Robotics and Automation (2005)
24. Zivkovic, Z., Kröse, B.: Part based people detection using 2D range data and images. In Proc. of the IEEE/RSJ Intl. Conf. Robots and Autonomous Systems (IROS) (2007)