Toward robust foveated wide field of view people detection

Zoran Zivkovic and Ben Kröse

ISLA lab, University of Amsterdam,
Kruislaan 403, 1098SJ Amsterdam, The Netherlands

Abstract. We present a foveated vision system for robust person detection in wide field of view. The system consists of an omidirectional camera for people detection in a wide field of view and a pan-tilt camera that can focus on a particular location. Combining the information from both cameras leads to improved people detection. The people detection is based on human body part detectors and a probabilistic model of the spatial arrangement of the parts. The representation is robust to partial occlusions, part detector false alarms and missed detections of body parts. We also show how to use the fact that the persons walk on a known ground plane to increase the efficiency and reliability of the detection.

1 Introduction

Person detection from images is a widely studied problem, e.g. [5, 8, 12, 9]. The part-based object representations, e.g. [11, 3], often lead to higher recognition performance when compared to algorithms considering a complex object as a whole. The second important advantage of the part-based approach is it relies on object parts and therefore it is much more robust to partial occlusions. People detection by detecting body parts was considered a number of times. Seemann et al. [8] use SIFT based part detectors to detect people but do not model part occlusions. Wu and Nevatia [12] describe the part occlusions but the occlusion probabilities and part positions are learned in a supervised manner.

In this paper we present part based people detection similar to [11, 4]. 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. We propose to use the Haar-like feature cascade classifier of Viola and Jones [10] to rapidly detect human body parts at various scales instead of the salient regions used before [11, 4]. In the experimental section we show that the probabilistic combination of part detections performs much better than each part separately and better than the Haar-like feature cascade classifier applied to the whole body. We also propose how to use the fact that the persons walk on a known floor plane to detect people more efficiently. Furthermore, in this paper we propose a vision system for reliable and robust people detection in wide field of view. The system consists of an omidirectional camera that delivers low-resolution wide field of view images and

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

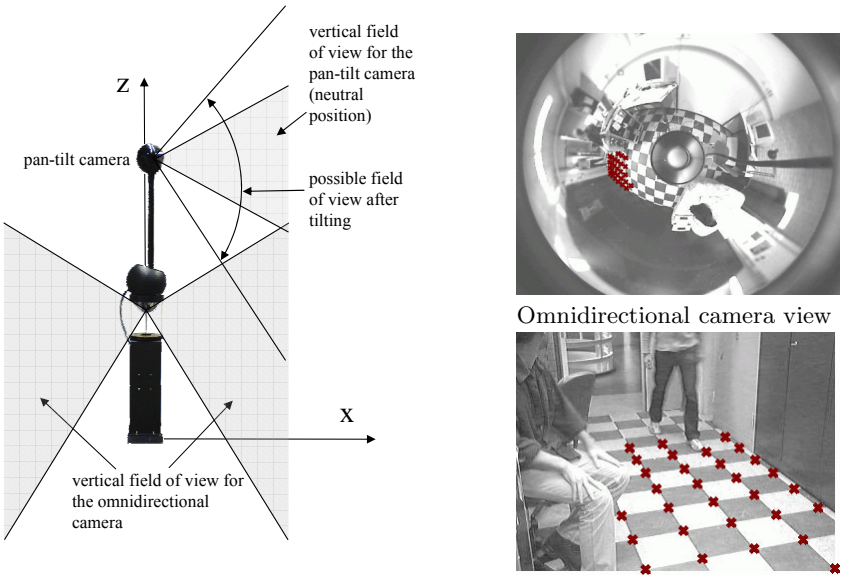


Fig. 1. The camera system construction. The omnicam has 360° horizontal field of view. The vertical viewing angle of the omnicam is 90° with 30° maximal upward angle. The pan-tilt unit can tilt 102° and pan 189°. We show example views from the two cameras on the left. We indicate the corner points corresponding to the floor tiles that are used for estimating the position of the cameras with respect to the floor plane.

a pan-tilt camera that can focus on a particular location. We show how our part-based model can be used to combine part detections from the two cameras.

This paper is organized as follows. We start with describing the camera system we use and the calibration of the cameras in Section 2. In Sections 3 and 4 we present part-based probabilistic model. The results from our experiments are in Section 5. Conclusions are in Section 6.

2 Description of the vision system

Our system consists of two cameras. A side view of the system is presented in Figure 1. The omnidirectional camera consists of a regular camera and a hyperbolic shape mirror in front of the camera. The camera delivers 1024×768 color images. The mirror is properly placed with respect to the camera lens and the camera with the mirror can be described using the standard central camera model. The omnicam images can be used for wide field of view people detection. However, because of the wide field of view the resolution is low for far away objects. Therefore we add a pan-tilt camera that can deliver higher resolution images for particular positions. This resambles biologically inspired systems [1, 2]. The system is intended to be used on a mobile robot that should interact with people, see Figure 4. The pan-tilt unit is placed on top of the

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

omnidirectional camera as depicted in Figure 1. When on our mobile robot, the pan-tilt camera is high enough to have good viewing position for observing human face. Additionally a moving pan-tilt unit can be used to make people aware of the current point of attention of the system[2]. Disadvantages of such construction are: a small part of the omnicam image is occluded by the pan-tilt camera data cable; and the pan-tilt unit can be shaking during the robot movements.

2.1 Calibration

First the intrinsic parameters of the both cameras are estimated. Next, we estimate the 3D position of the both cameras with respect to the floor plane, see Figure 4. This is achieved by identifying the same known 3D points in both cameras. In our case we used the corners of the floor tiles as shown in Figure 1. The pose of the both cameras is estimated then by minimizing the reprojection errors of the selected points [6]. The camera pose for the pan-tilt unit is estimated for the "maximal tilt down" and "neutral pan" camera position and we rely on the pan-tilt angles reported by the camera pan-tilt mechanism.

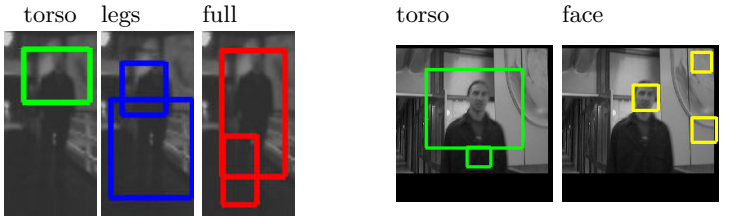
3 Probabilistic part-based person model

The human body will be represented as a collection of P body parts. The 2D image position of the p-th part will be denoted by $\mathbf{x}_p = (x_p, y_p)$. We will use the Gaussian distribution for the arrangement of the body parts:

$$p_{shape}(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu, \Sigma) \tag{1}$$

where $\mathbf{x} = \left(\mathbf{x}_1 \ ... \ \mathbf{x}_P\right)$ is a $2P$ long vector containing all the 2D part positions, μ is the mean and Σ 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 [3].

3.1 Part detection



Part detections – omnicam Part detections – pan-tilt camera

Fig. 2. Example body part detections with some false detections.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

We use a set of Haar-like-feature classifiers to detect various human body parts [10]. In this paper the classifiers are trained on face, upper body, lower body and full body images. The upper body, lower body and full body detectors are applied to the omnicam images and in pan tilt images we detect face and upper body. This gives in total $P = 5$ parts, see Figure 2. Let N_p denote the number of detections of part p. The positions of all detected parts can be summarized in a data structure:

$$\mathcal{X} = \begin{pmatrix} \mathbf{x}_{1,1} & \mathbf{x}_{1,2} & \dots & \mathbf{x}_{1,N_1} & \\ \mathbf{x}_{2,1} & \mathbf{x}_{2,2} & \dots & \dots & \mathbf{x}_{2,N_2} \\ \dots & \dots & \dots & \dots & \dots \\ \mathbf{x}_{P,1} & \dots & \mathbf{x}_{P,N_P} & & \end{pmatrix} \tag{2}$$

with one row per part and where each row contains information about the detections of the corresponding body part. The element $\mathbf{x}_{p,j}$ contains the 2D image position $(x_{p,j}, y_{p,j})$ 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 is not detected.

3.2 Missing detections and clutter

From an image we will extract a collection of parts $\mathcal{X}$ some of which might be false detections. To indicate which detections correspond to the object we will use vector $\mathbf{h}$ with element $h_p = j$, $j > 0$, indicating that the j-th detection of the p-th part $\mathbf{x}_{p,j}$ belongs to the object. Other detections of that part are false detections. Given $\mathbf{h}$ the shape of the object is composed of the corresponding detections $\mathbf{x} = \left(\mathbf{x}_{1,h_1} \dots \mathbf{x}_{P,h_P}\right)$. The detections that belong to the background clutter are denoted as $\mathbf{x}^{bg}$.

It is also possible that all detections are false detections or the part was not detected at all. We use $h_i = 0$ to indicate this. These parts are considered as missing data. We will denote the set of missing parts as $\mathbf{x}^m$ and the set of observed parts as $\mathbf{x}^o$. The probabilistic model of the arrangement of the body parts (3) will be written as:

$$p_{shape}(\mathbf{x}) = p_{shape}(\mathbf{x}^o, \mathbf{x}^m) \tag{3}$$

For a collection of detected parts we do not know $\mathbf{h}$ and it is also treated as missing (hidden) data. We will call $\mathbf{h}$ the 'hypothesis' vector. If there are N_p detections for part p, the number of possible hypotheses is $\prod_p^P (N_p + 1)$.

3.3 Probabilistic model

A probabilistic model that considers the possibility of part detector false alarms and missed detections of body parts of a person can be written as:

$$p(\mathcal{X}, \mathbf{x}^m, \mathbf{h}) = p(\mathcal{X}, \mathbf{x}^m | \mathbf{h}) p(\mathbf{h}) \tag{4}$$

where $\mathbf{x}^m$ are the missing data determined by $\mathbf{h}$.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

In order to define $p(\mathbf{h})$ we will introduce two auxiliary variables $\mathbf{b}$ and $\mathbf{n}$. The variable $\mathbf{b} = sign(\mathbf{h})$ 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. We can now write the joint distribution (4) as:

$$p(\mathcal{X}, \mathbf{x}^m, \mathbf{h}, \mathbf{n}, \mathbf{b}) = p(\mathcal{X}, \mathbf{x}^m|\mathbf{h})p(\mathbf{h}|\mathbf{n}, \mathbf{b})p(\mathbf{n})p(\mathbf{b}) \tag{5}$$

where we add the two auxiliary variables $\mathbf{b}$ and $\mathbf{n}$ and assume independence between them. Furthermore, we have:

$$p(\mathcal{X}, \mathbf{x}^m|\mathbf{h}) = p_{shape}(\mathbf{x}^o, \mathbf{x}^m)p_{bg}(\mathbf{x}^{bg}) \tag{6}$$

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 $\mathbf{h}$ and the $p_{bg}(\mathbf{x}^{bg})$ describes the distribution of the false detections. We will assume uniform density for the false detections:

$$p_{bg}(\mathbf{x}^{bg}) = \prod_{p=1}^{P}(1/A)^{n_p}. \tag{7}$$

where A is the total image area and $n_p \leq N_p$ is the element from the vector $\mathbf{n}$.

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 $p(\mathbf{b})$.

We assume here that the background part detections $\mathbf{n}$ are independent of each other and modelled using Poisson distribution with mean M_p [11]. Different M_p-s for different parts admit different detector statistics. The Poisson parameter will be denoted by vector $\mathbf{M} = \begin{pmatrix} M_1 & ... & M_P \end{pmatrix}$.

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

$$p(\mathbf{h}|\mathbf{n}, \mathbf{b}) = \begin{cases} 1/|\mathcal{H}(\mathbf{b}, \mathbf{n})| & \text{if } \mathbf{h}\epsilon\mathcal{H}(\mathbf{b}, \mathbf{n}), \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

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.

3.4 Learning model parameters

The density distribution (5) will have the following set of parameters $\Omega = \{\mu, \Sigma, p(\mathbf{b}), \mathbf{M}\}$. Therefore we can write the distribution (5) also as:

$$p(\mathcal{X}, \mathbf{x}^m, \mathbf{h}) = p(\mathcal{X}, \mathbf{x}^m, \mathbf{h}|\Omega) \tag{9}$$

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

$$p(\mathcal{X}|\Omega) = \sum_{\text{all possible } \mathbf{h}} \int_{\mathbf{x}^m} p(\mathcal{X}, \mathbf{x}^m, \mathbf{h}|\Omega). \tag{10}$$

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

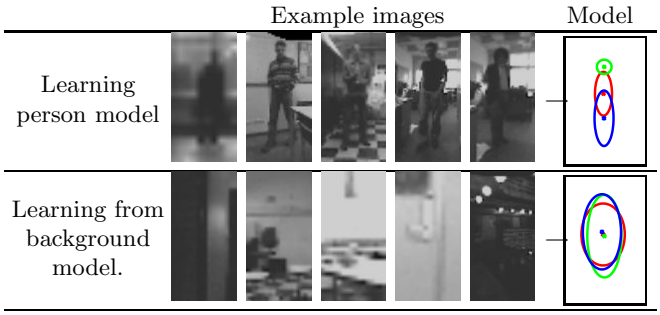


Fig. 3. Example images from the data set used to train the probabilistic part arrangement model. For each part we present its mean position contained in the parameter μ. The ellipse represents the 1-sigma uncertainty of the part position as described by the diagonal elements of the covariance matrix Σ. Here green color represents head, blue are legs and red is the full body detector.

We use Gaussian distribution to describe the arrangement of the body parts. 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 we start from a set of L aligned images of persons. The part detectors are applied to each image. The collection of detected parts for i-th image will be denoted as $\mathcal{X}_i$. The maximum likelihood estimate of the parameters Ω is computed by maximizing the likelihood of the data: $\prod_i^L p(\mathcal{X}_i|\Omega)$ Once we have the part detectors, the part arrangement parameters are estimated using expectation maximization algorithm from a set of unlabelled images [11].

3.5 Detection

Let us denote the maximum likelihood parameters learned from a set of images of persons as Ω_{person}, see Figure 3. The parameters of the model can be learned also for a set of random images of the background from the environment. These parameters will be denoted as Ω_{bg}. We are now presented with a new image and extracted the set of detected parts $\mathcal{X}$. The image is either an image of a person or some background image:

$$p(\mathcal{X}) = p(\mathcal{X}|Person)p(Person) + p(\mathcal{X}|BG)p(BG) \tag{11}$$

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

$$p(Person|\mathcal{X}) = \frac{p(\mathcal{X}|Person)p(Person)}{p(\mathcal{X})} \approx \frac{p(\mathcal{X}|\Omega_{person})p(Person)}{p(\mathcal{X}|\Omega_{person})p(Person) + p(\mathcal{X}|\Omega_{bg})p(BG)} \tag{12}$$

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

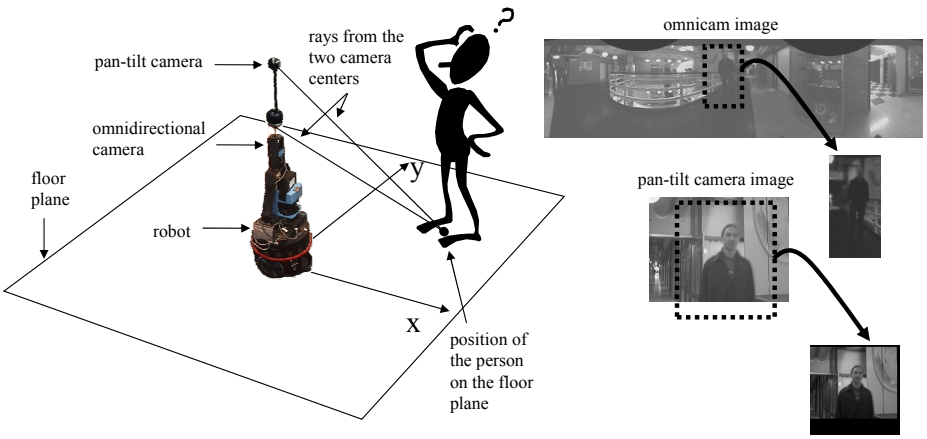


Fig. 4. Schematic representation of our Nomad robot with the camera system. Given a floor plane position of the person, the regions of interest can be extracted from both camera images as illustrated on the right.

The last step above is an approximation since we use the maximum likelihood estimates for the model parameters Ω_{person} and Ω_{bg} instead of integrating over all possible parameter values. Calculating $p(\mathcal{X}|\Omega)$ is done using (10).

4 People detection with floor plane constraint

In practice we do not know where the person is in a new image and at which scale. Therefore, person detection would standardly involve scanning the image across many possible image positions and scales. This is computationally expensive and often can be done more efficiently using the floor plane constraint.

4.1 Floor plane constraint

We assume that people walk over a flat ground floor surface. This is in general true for most man-made environments. Therefore we will define a set of possible 2D positions $\mathcal{T}_t$ of the person on the floor plane. In our experiments we consider a $10m \times 10m$ square area around the robot on the ground floor and a grid of possible positions at every $10cm$. This gives a total of 10000 possible ground floor points $\mathcal{T}_t$. Then, instead of scanning an image across all possible image positions and scales we will just scan all predefined ground floor points $\mathcal{T}_t$.

We used data from the National Center for Health Statistics (*www.cdc.gov/nchs/*). For adult humans, the mean height is 1.7m with a standard deviation of 0.085m. We define maximal height of human to be mean plus three standard deviations and the width to be 1/2 of the height. Using these dimensions and having a calibrated camera, each $\mathcal{T}_t$ defines a rectangle region of interest in an image, see Figure 4. Keeping only the parts within the possible regions of interest greatly reduces the number of parts that need to be considered, Figure 5.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

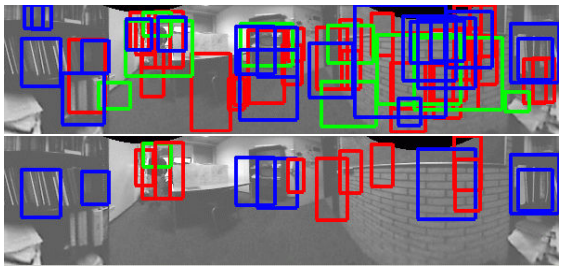


Fig. 5. Body part detection (top). With the floor plane contraint (below).

4.2 People detection

Each 2D floor position $\mathcal{T}_t$ has a set of parts $\mathcal{X}$ within the corresponding region of interest of the image. The likelihood of the parts $p(\mathcal{X}, \mathbf{x}^m, \mathbf{h}|\Omega_{person}, \mathcal{T}_t)$ for the given position $\mathcal{T}_t$ is computed by (5) and we can write:

$$p(\mathcal{X}, \mathbf{x}^m, \mathbf{h}, \mathcal{T}_t|\Omega_{person}) = p(\mathcal{X}, \mathbf{x}^m, \mathbf{h}|\Omega_{person}, \mathcal{T}_t)p(\mathcal{T}_t) \tag{13}$$

where $p(\mathcal{T}_t)$ is some prior distribution on the possible person locations on the floor plane. The region of interest image is either an image of a person or some background image and we can write:

$$p(Person|\mathcal{X}, \mathcal{T}_t) \approx \frac{p(\mathcal{X}, \mathcal{T}_t|\Omega_{person})p(Person)}{p(\mathcal{X}, \mathcal{T}_t|\Omega_{person})p(Person) + p(\mathcal{X}, \mathcal{T}_t|\Omega_{bg})p(BG)} \tag{14}$$

This is an approximation since we use the maximum likelihood estimates for the model parameters Ω_{person} and Ω_{bg} instead of integrating over all possible parameter values. Calculating $p(\mathcal{X}, \mathcal{T}_t|\Omega)$ for a given $\mathcal{T}_t$ is done using (10). We assume a priori probabilities to be equal $p(Person) = p(BG) = 0.5$ and decide that there is a person at position $\mathcal{T}_t$ if $p(Person|\mathcal{X}, \mathcal{T}_t) > 0.5$.

Since $p(Person|\mathcal{X}, \mathcal{T}_t)$ is computed at a dense grid of ground points, it often happens that $p(Person|\mathcal{X}, \mathcal{T}_t)$ has a large value for a number of ground points around the position where the person actually is. Therefore the persons are detected as the local maxima of the $p(Person|\mathcal{X}, \mathcal{T}_t)$ above the threshold value.

5 Experiments

5.1 Recognition results

The Haar-like-feature based part detectors we used in our experiments were trained on the MIT pedestrian dataset [7] and are available in the Intel OpenCV library. We also used the face detector provided in the Intel library.

For training our part-based model we made another data set of 400 low resolution images of people cut out of panoramic images and from pan-tilt camera images. The images are obtained by our robot by driving around our office

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

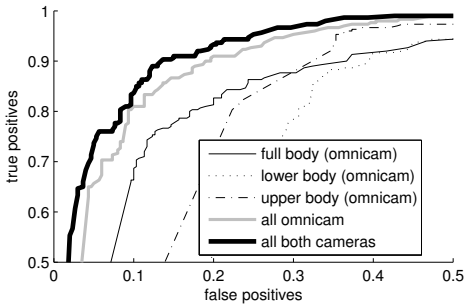


Fig. 6. Recognition Receiver Operator Curves.

space. We used 200 images for learning the model parameters and the other 200 images for testing the recognition results. We randomly cut out also a set of 1200 background images. The results in Figure 6 show that large improvements can be achieved if the part detectors are combined. Combining the images from both cameras further improves the results.

5.2 Recognition from a moving robot

The algorithm presented in the previous section is implemented on our robot. In Figure 7 we present a few panoramic images with the detection results. The images were obtained at various places around our building. The detection accuracy is much better than the single part detectors.

The first stage of the algorithm where the body parts are detected is the most computationally expensive. Running three Haar-like-feature based part detectors on a 600×150 panoramic image takes on average 400ms on a 2GHz single processor computer. 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 can not fit into the image anymore. However, the floor constraint can heavily reduce the number of positions and part sizes to search for and the time can be cut down to around 100ms. The two detectors (upper body and face) for the 320×200 pan-tilt camera image take 80ms. Once the parts are detected, detecting persons using our model takes around 20ms.

6 Conclusions and further work

We presented a foveated vision system for robust person detection. An omidirectional camera is combined with a pan-tilt camera that can focus on a particular location. We used Haar-feature based cascade classifiers to detect different human body parts: upper body, lower body, face and full body. We present a principled probabilistic representation that combines the part detections and can achieve person detection robust to partial occlusions, part detector false alarms and missed detections of body parts. The recognition results greatly outperform each of the single Haar-feature based cascade classifiers. Combining the information from both cameras leads to more reliable people detection.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

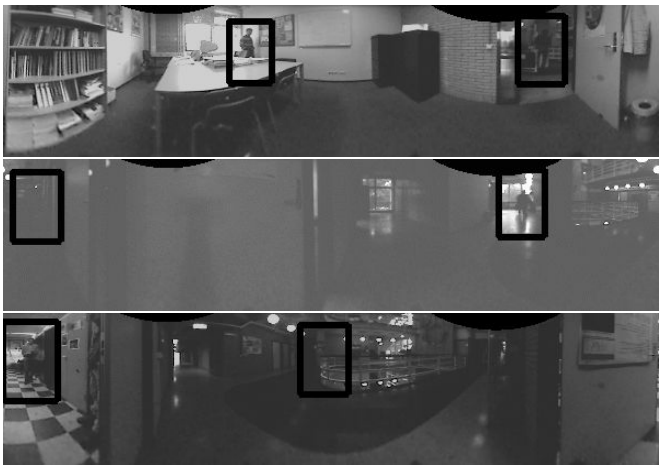


Fig. 7. Example people detections in panoramic images recorded from a moving robot. One false positive detection is in the last image.

References

1. D.H. Ballard. Animate vision. *Artificial Intelligence*, 48:57–86, 1991.
2. C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati. Active vision systems for sociable robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 31(5):443–453, 2001.
3. P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *Intl. Journal of Computer Vision*, 61(1):55–79, 2005.
4. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.
5. D.M. Gavrila and V. Philomin. Real-time object detection for smart vehicles. *In Proc. of the Intl. Conf. on Computer Vision*, 1999.
6. R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.
7. H. Kruppa, M. Castrillon-Santana, and B. Schiele. 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.
8. E. Seemann, B. Leibe, K. Mikolajczyk, and B. Schiele. An evaluation of local shape-based features for pedestrian detection. *In Proc. of the British Machine Vision Conference*, 2005.
9. S.Munder and D. M. Gavrila. An experimental study on pedestrian classification. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(11):1863–1868, 2006.
10. P.A. Viola and M.J. Jones. Rapid object detection using a boosted cascade of simple features. *In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.
11. M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. *In Proc. of the European Conf. on Computer Vision*, 2000.
12. B. Wu and R. Nevatia. 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.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de