# People Detection by a Mobile Robot Using Stereo Vision in Dynamic Indoor Environments

José Alberto Méndez-Polanco, Angélica Muñoz-Meléndez,
and Eduardo F. Morales

National Institue of Astrophysics, Optics and Electronics
{polanco,munoz,emorales}@inaoep.mx

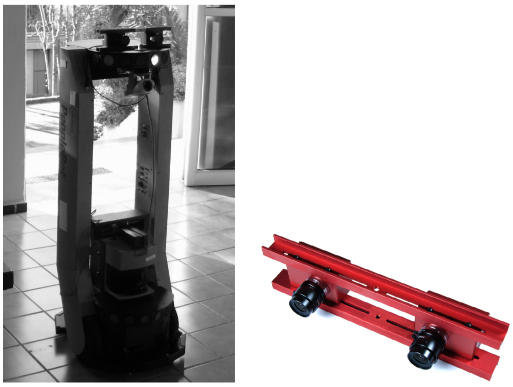
**Abstract.** People detection and tracking is a key issue for social robot design and effective human robot interaction. This paper addresses the problem of detecting people with a mobile robot using a stereo camera. People detection using mobile robots is a difficult task because in real world scenarios it is common to find: unpredictable motion of people, dynamic environments, and different degrees of human body occlusion. Additionally, we cannot expect people to cooperate with the robot to perform its task. In our people detection method, first, an object segmentation method that uses the distance information provided by a stereo camera is used to separate people from the background. The segmentation method proposed in this work takes into account human body proportions to segment people and provides a first estimation of people location. After segmentation, an adaptive contour people model based on people distance to the robot is used to calculate a probability of detecting people. Finally, people are detected merging the probabilities of the contour people model and by evaluating evidence over time by applying a Bayesian scheme. We present experiments on detection of standing and sitting people, as well as people in frontal and side view with a mobile robot in real world scenarios.

## 1 Introduction

Traditionally, autonomous robots have been developed for applications requiring little interaction with humans, such as sweeping minefields, exploring and mapping unknown environments, inspecting oil wells or exploring other planets. However, in recent years significant progress has been achieved in the field of service robots, whose goal is to perform tasks in populated environments, such as hospitals, offices, department stores and museums. In these places, service robots are expected to perform some useful tasks such as helping elderly people in their home, serving as hosts and guides in museums or shopping malls, surveillance tasks, childcare, etc [1,2,3,4,5,6]. In this context, people detection by a mobile robot is important because it can help to improve the human robot interaction, perform safety path planning and navigation to avoid collisions with people, search lost people, recognize gestures and activities, follow people, and so on.

A. Hernández Aguirre et al. (Eds.): MICAI 2009, LNAI 5845, pp. 349–359, 2009.

In this work, we propose to detect people applying a mobile robot using a semi-elliptical contour model of people. The main difference between our proposed model and previously proposed models is that our model takes into account the distance between the robot and people to try to get the best fit model. Therefore, the contour model detects people without assuming that the person is facing the robot. This difference is important because we cannot expect people to cooperate with the robot to perform its task. The contour model is used to get a first estimate of the position of a person relative to the robot. However, instead of detecting people in a single frame, we apply a spatio-temporal detection scheme merging the probabilities of the contour people model over time by applying a Bayesian scheme. The idea is that people who are detected by the robot at different instants of time have a higher probability of detection. Our experimental platform is an ActivMedia PeopleBot mobile robot equipped with a stereo camera (See Figure 1).



(a) (b)

**Fig. 1.** (a) ActivMedia PeopleBot equipped with a stereo camera. (b) Stereo camera.

The paper is organized as follows: In section 2 we present related works. Section 3 describes our segmentation method based on the distance to detected objects. Section 4 introduces the adaptive contour model to get a first estimation of the people position relative to the robot. Section 5 presents the detection method using a spatio-temporal approach. Section 6 shows the principal experiments and results. Finally, Section 7 presents conclusions and future research directions.

## 2 Related Work

Depending on the specific application that integrates people detection and identification there are two princ460,13ipal approaches that can be applied over images acquired: whole human body detection [7,8,9] and part-based body detection [10,11], for instance, with single, stereo CCD or thermal cameras.

The advantages of whole human body detection approaches are the compact representation of the human body models such as human silhouettes [12,13,14]. The main drawbacks of these approaches are the sensitivity to object occlusion and cluttered environments as well as the need of robust object segmentation methods. Concerning part-based body detection approaches, the main advantage is their reliability to deal with cluttered environments and object occlusion because of their independence for whole human body parts detection. These approaches do not rely on the segmentation of the whole body silhouettes from the background. Part-based body detection approaches, in general, aim to detect certain human body parts such as face, arms, hands, legs or torso. Different cues are used to detect body parts, such as laser range- finder readings to detect legs [15,16] or skin color to detect hands, arms and faces [17,18,19].

Although people recognition using static cameras has been a major interest in computer vision [18,20,21], some of these works cannot be directly applied to a mobile robot which has to deal with moving people in cluttered environments. There are four main problems that need to be solved for people detection with mobile robots: real time response, robust object segmentation, incomplete or unreliable information cues, and the integration of spatio-temporal information. Below we briefly explain these problems.

1. **Real time people detection.** Service robots must operate robustly in real time in order to perform their tasks in the real world and achieve appropriate human robot interaction. There are some real time object detection approaches which use SIFT features [23] or the Viola and Jones algorithm for face detection [22] that have been applied for people detection purposes. However people detection is more complicated due to possible poses of the human body and the unpredictable motion of arms and legs.
2. **Object segmentation.** General object segmentation approaches assume a static background and a static camera [18,20,21]. In this way, object segmentation can be achieved subtracting the current image from the background reference resulting in a new image with the possible people detected. In the case of mobile robots, to have a background reference for each possible robot location is not feasible, so traditional segmentation is no longer applicable.
3. **Incomplete or unreliable information cues.** In the context of service robots both, people and robots move within dynamic environments. For this reason, information cues necessary for people recognition such as faces or body parts are not always available. In the case of face detection, faces are not always perceivable by the camera of a mobile robot. Concerning legs detector, their main drawbacks are the number of false positives which may occur, for instance, in situations where people′s legs are indistinguishable from the legs of tables or chairs. As far as skin color detection approaches are concerned, the principal problem is that, in cluttered environments, there are typically many objects similar in color to human skin and people are not always facing the robot.
4. **Integration of spatio-temporal information.** During navigation, robots perceive the same objects from different locations at different periods of

time. A video streaming can be used to improve the recognition rate using evidence from several frames. This is conceptually different from most object recognition algorithms in computer vision where observations are considered to be independent.

## 3 Distance Based Image Segmentation

In this paper, we propose an image segmentation method based on distance. In order to achieve the distance information, a camera stereo is used to calculate the disparity between two images (left and right images). The idea of using a disparity image is that objects closer to the camera have a greater disparity than objects further to the camera. Therefore, the distance to the objects can be calculated with an adequate calibration. An example of a disparity image calculated from the images provided by a stereo camera is shown in Figure 2.

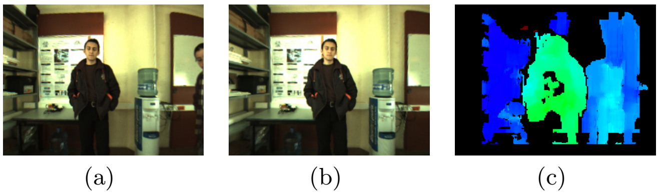


**Fig. 2.** Stereo images. (a) Left image. (b) Right image. (c) Disparity image. The color of each pixel in the disparity image indicates the disparity between the left and the right image where: clear colors indicate high disparity, dark colors indicate low disparity and black pixels indicate no disparity information.

Once the distance to the objects has been calculated, we scan one of the stereo camera images to segment objects based on the previous distance information. The idea is to form clusters of pixels with similar distances to the robot ($Z$ coordinate) and, at the same time, near each other in the image (($X$,$Y$) coordinate). The clusters are defined as follows:

$$C_k = \{\mu_X^k, \mu_Y^k, \mu_Z^k, \sigma_X^k, \sigma_Y^k, \sigma_Z^k, \rho_k\}, k = 1...N \tag{1}$$

where $\mu_X^k$, $\mu_Y^k$ and $\mu_Z^k$ are the mean of the $X, Y, Z$ coordinates of the pixels within the cluster $k$, $\sigma_X^k$, $\sigma_Y^k$ and $\sigma_Z^k$ are the variances of the $X, Y, Z$ coordinates of the pixels within the cluster $k$, $\rho_k$ is a vector containing the coordinates of the pixels in the cluster $k$, and $N$ defines the maximum number of clusters.

The means and the variances of each cluster are calculated from the data provided by the disparity image. The image is then scanned and for each pixel we verify if $X, Y, Z$ coordinates are near to one of the clusters. The segmentation consists of three main steps for each pixel in the image as described below:

1. Calculate the distances of the pixel to the clusters based on the means of the clusters:

$$d_X^k = (X - \mu_X^k)^2, d_Y^k = (Y - \mu_Y^k)^2, d_Z^k = (Z - \mu_Z^k)^2 \tag{2}$$

2. If the cluster $k$ has the lowest distance and $d^k_X < th_X$ and $d^k_Y < th_Y$ and $d^k_Z < th_Z$ then assign the pixel to the cluster $k$ and update the means and the variances of the cluster. In this case, $th_X$, $th_Y$ and $th_Z$ are thresholds previously defined.
3. If there are no clusters closed to the pixel create a new cluster and assign the pixel to this cluster.

We use three different thresholds to perform the segmentation to exploit the fact that, in most cases, the height of people in the images is larger than the width. Therefore, we define $th_X < th_Y$ in order to segment objects with the height greater than the width, as in the case of standing and most sitting human bodies. Further more, after segmentation, clusters with irregular proportions ($\mu_X > \mu_Y$) are eliminated to take into account only objects with human proportions. In Figure 3 two views of the coordinate system for the images shown in Figure 2 are illustrated.

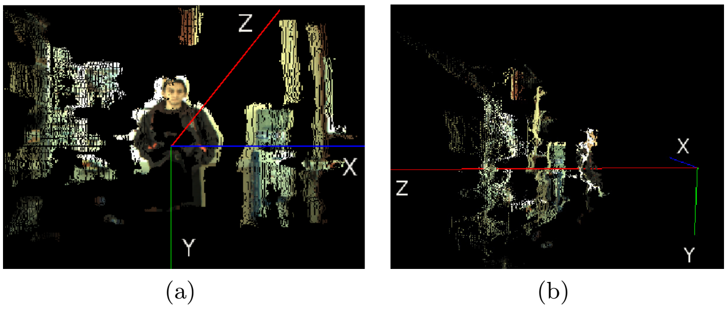


**Fig. 3.** Coordinate system for the image shown in Figure 2. (a) Front view. (b) Top view.

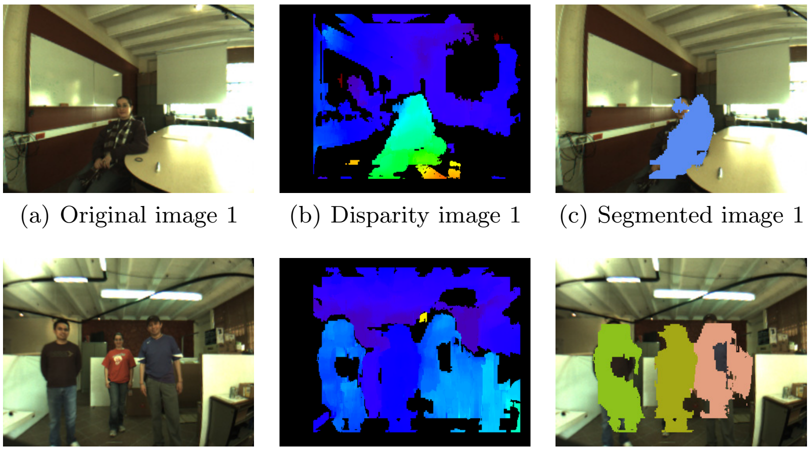
Figure 4 shows examples of the object segmentation based on distance using the method described in this section. This example illustrates the reliability of our method to segment simultaneously one or more people.

## 4 Adaptive Semi-elliptical Contour Model

The segmentation method provides different regions where there are possible people. The next step is to determine which of those regions effectively contain people and which do not. In order to do that, we apply a semi-elliptical contour model illustrated in Figure 5, similar to the model used in [25]. The semi-elliptical contour model consists of two semi-ellipses describing the torso and the human head. The contour model is represented with an 8- dimensional state vector:

$$d^t_{body} = (x_T, y_T, w_T, h_T, x_H, y_H, w_H, h_H) \quad (3)$$

where $(x_T, y_T)$ is the mid-point of the ellipse describing the torso with width $w_T$ and height $h_T$, and $(x_H, y_H)$ is the mid-point of the ellipse describing the head with width $w_H$ and height $h_H$.



(a) Original image 1 (b) Disparity image 1 (c) Segmented image 1

(d) Original image 2 (e) Disparity image 2 (f) Segmented image 2

**Fig. 4.** Segmentation method. These examples illustrate how our segment method is able to segment one or multiple people. (a) and (d) are the original images, (b) and (e) are the disparity images, and (c) and (f) are the result of our segmentation process.

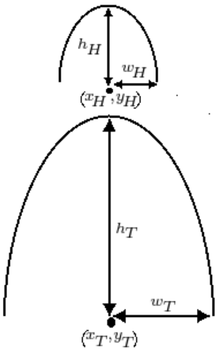


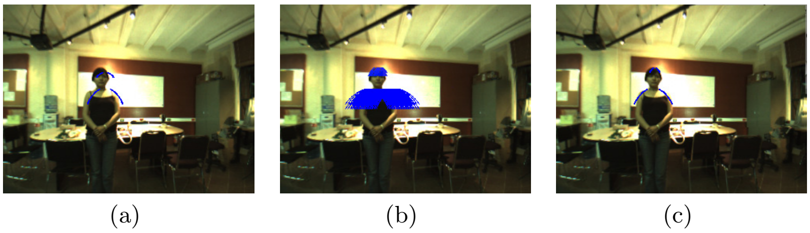
**Fig. 5.** Semi-elliptical contour model of people similar to the model used in [25]

For each region obtained by the segmentation method an elliptic model is fitted setting the mid-point of the torso ellipse to the center of the region. Since people width varies with distance, we determine the dimension of the elliptic model using a table containing the means of the width of torso and head for five different people at different distances to the robot. This constraint avoids considering regions whose dimensions are incompatible with the dimensions of people at specific distances. At the same time, this constraint enables the robot to achieve a better fit for the semi-ellipsis describing people.

As different people have usually different width, we adjust the contour people model by varying the mid-point of the torso as well as its width and its height. To determine which variation has the best fit, a probability of fitting for each variation is calculated. The idea is to evaluate the fitting probability as the probability of the parameters of the contour people model given a person $p_i$, that we denote as $P(d^v_{body}|p_i)$. The probability is calculated as follows:

$$P(d^v_{body}|p_i) = argmax \frac{(N_f|d^v_{body})}{N_T} \tag{4}$$

where $N_f$ is the number of points in the ellipse that fit with an edge of the image given the parameters of the model $d^t_{body}$, $v$ denotes the different variations of the model and $N_T$ is the total number of points in the contour model. The edges of the image are calculated applying the Canny edge detector [24]. The adjustment process for the contour people model is illustrated in Figure 6.



(a) (b) (c)

**Fig. 6.** Adjustment process for the contour people model. (a) Original application of the contour model. (b) adjustment process by varying the model parameters. (c) Final application of the contour model.

## 5 People Detection and Tracking

To detect people, we obtain first the regions of interest using the method described in Section 3. After that, we evaluate the probability of detection using the contour people model described in Section 4. At this point we can determine, with certain probability, the presence of people in the image. In this work, we improve over traditional-frame based detector incorporating cumulative evidence from several frames using a Bayesian scheme.

Once a person $p_i$ has been detected at time $t$, we proceed to search if that person was previously detected at time $t-1$ calculating the Euclidean distances from the position of the current person to the positions of the previously detected people. If the distance between two people is less than a threshold, then we consider these people to be the same. Once two people have been associated we proceed to calculate the probability of a person $p_i$ at time $t$ denoted as $P(p^t_i)$ as follows:

$$P(p^t_i) = \frac{P(d^t_{body}|p_i)P(p^{t-1}_i)}{(P(d^t_{body}|p^t_i)P(p^{t-1}_i) + (P(d_{body}| \sim p^t_i)P(\sim p^{t-1}_i)} \tag{5}$$

where $P(d^t_{body}|p_i)$ is the probability of fitting the contour people model at time $t$ which is calculated applying equation 4. $P(p^{t-1}_i)$ is the probability of the person

$p_i$ at time $t-1$. $P(d_{body}| \sim p_i^t)$ is the probability of not fitting of the contour people model at time $t$ and $P(\sim p_i^{t-1})$ is the probability that the person $p_i$ has not been detected at time $t-1$.

## 6 Experimental Results

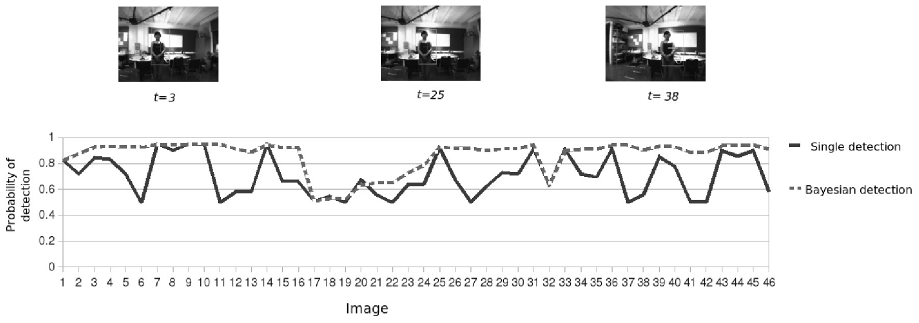
We tested our people detection method using a mobile robot equipped with a stereo camera. The experiments were performed in a dynamic indoor environment under different illumination conditions and with sitting and standing people placed at different distances from the robot (1 to 5 $m$). The number of people varies from 1 to 3 people. Due to the fact that we do not use the face as a cue to detect people our method can detect people facing or not the robot. Figure 7 compares the performance of our people detection method using a single frame detection scheme against our people detection method using evidence from several frames applying a Bayesian scheme. Figure 8 shows how our people detection method is able to detect multiple people and track them over time. We consider a person as detected if $P(p_i) > 0.8$. We calculated the detection rate $DR$ as follows:

$$DR = \frac{N_D}{N_T} \tag{6}$$

where $N_D$ is the number of frames where people were detected with $P(p_i) > 0.8$, and $N_T$ is the total number of frames analysed.

Our detection method has a $CR$ of 89.1% using $P(p_i) > 0.8$ and 96% if we use $P(p_i) > 0.6$. Table 1 presents the detection rate of our method compared with the classification rate reported in [25] which presents two different approaches to detect people using a thermal camera and a semi-elliptic contour model.

In Figure 9 one can see the results of different experiments on detection of standing and sitting people, as well as people in frontal and side view with a mobile robot in real world scenarios using our proposed people detection method are shown.

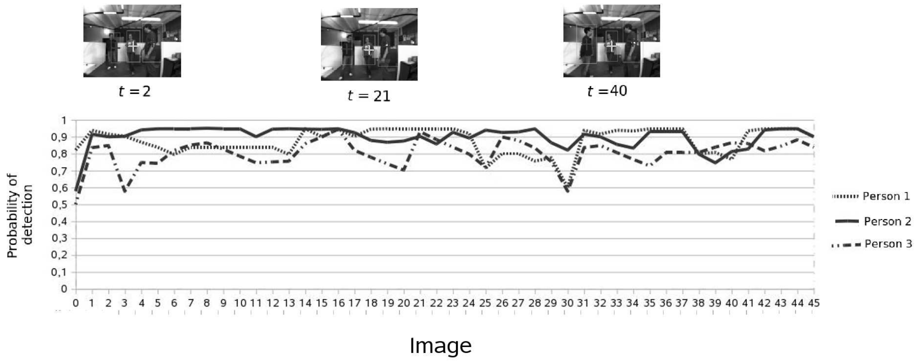


(a)

**Fig. 7.** People detection performance. The Bayesian people detection outperforms the single frame detection. At the top of the chart we show images at different periods of time showing the output of our people detection method.

**Table 1.** Comparation with other works

| **Method** | **Classification rate** |
|---|---|
| Contour [25] | 88.9 |
| Combination with grey features [25] | 90.0 |
| Our Method (Spatio-temporal) | 96.0 |



(a)

**Fig. 8.** Multiple people detection performance

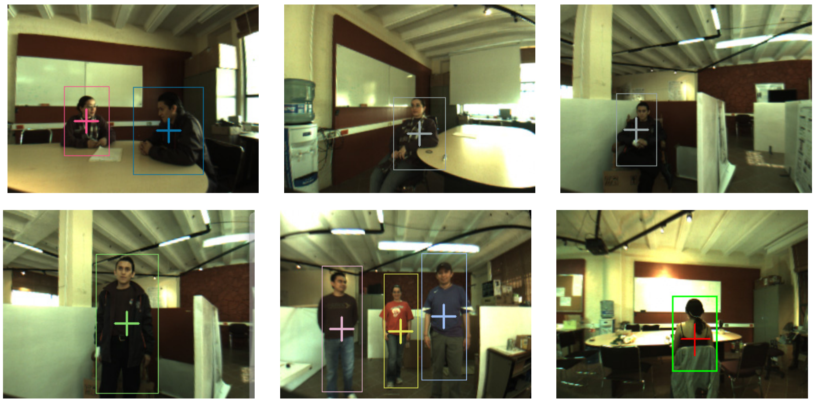


**Fig. 9.** Experiments performed using a mobile robot in dynamic indoor environments

## 7 Conclusion and Future Work

This paper addressed the problem of detecting people with a mobile robot using a stereo camera. The proposed segmentation method takes into account human body proportions to segment people and to provide a first estimation of people location. We presented an adaptive contour people model based on people distance to the robot. To detect people, we merged the probabilities of the contour people model over time by applying a Bayesian scheme. According to the experiments evidence, we show that our method is able to segment and detect standing and sitting people in both front and lateral views. Neither previous visual model of the environment nor mandatory facing pose of people is involved in our method. The future research directions of this work are fusing information from diverse cues such as skin, clothes and body parts to reduce the number of false positives, and the incorporation of a semantic map with *a priori* information about the probable location of people. Concerning people tracking a simple Euclidean distance based tracker has been used at this stage of our research. However, the integration of appearance model, motion model and a Kalman filter are considered in the near future to improve the people tracking process.

## References

1. Burgard, W., Cremers, A., Fox, D., Hähnel, D., Lakemeyer, G., Schulz, D., Steiner, W., Thrun, S.: Experiences with an Interactive Museum Tour-guide Robot. Artificial Intelligence, 3–55 (1999)
2. Osada, J., Ohnaka, S., Sato, M.: The Scenario and Design Process of Childcare Robot, PaPeRo, pp. 80–86 (2006)
3. Gharpure, C.P., Kulyukin, V.A.: Robot-assisted Shopping for the Blind: Issues in Spatial Cognition and Product Selection. Intelligent Service Robotics 1(3), 237–251 (2008)
4. Forlizzi, J., DiSalvo, C.: Service robots in the domestic environment: a study of the roomba vacuum in the home. In: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction, Salt Lake City, Utah, USA, pp. 258–265 (2006)
5. Montemerlo, M., Pineau, J., Roy, N., Thrun, S., Verma, V.: Experiences with a Mobile Robotic Guide for the Elderly. In: Proceedings of the 18th national conference on Artificial intelligence, Edmonton, Alberta, Canada, pp. 587–592 (2002)
6. Gockley, R., Forlizzi, J., Simmons, R.: Interactions with a Moody Robot. In: Interactions with a Moody Robot, pp. 186–193. ACM, New York (2006)
7. Malagón-Borja, L., Fuentes, O.: Object Detection Using Image Reconstruction with PCA. Image Vision Computing 27, 2–9 (2009)
8. Cielniak, G., Duckett, T.: People Recognition by Mobile Robots. Journal of Intelligent and Fuzzy Systems: Applications in Engineering and Technology 15(1), 21–27 (2004)
9. Davis-James, W., Sharma, V.: Robust Detection of People in Thermal Imagery. In: Proceedings of 17th International Conference on the Pattern Recognition ICPR 2004, pp. 713–716 (2004)
10. Müller, S., Schaffernicht, E., Scheidig, A., Hans-Joachim, B., Gross-Horst, M.: Are You Still Following Me? In: Proceedings of the 3rd European Conference on Mobile Robots ECMR, Germany, pp. 211–216 (2007)
11. Darrell, T.J., Gordon, G., Harville, M., Iselin-Woodfill, J.: Integrated Person Tracking Using Stereo, Color, and Pattern Detection. International Journal of Computer Vision, 175–185 (2000)
12. Vinay, S., James, W.D.: Extraction of Person Silhouettes from Surveillance Imagery using MRFs. In: Proceedings of the Eighth IEEE Workshop on Applications of Computer Vision, p. 33 (2007)

13. Ahn, J.-H., Byun, H.: Human Silhouette Extraction Method Using Region Based Background Subtraction. LNCS, pp. 412–420. Springer, Berlin (2007)
14. Lee, L., Dalley, G., Tieu, K.: Learning Pedestrian Models for Silhouette Refinement. In: Proceedings of the 19th IEEE International Conference on Computer Vision, Nice, France, pp. 663–670 (2003)
15. Schaffernicht, E., Martin, C., Scheidig, A., Gross, H.-M.: A Probabilistic Multimodal Sensor Aggregation Scheme Applied for a Mobile Robot. In: Proceedings of the 28th German Conference on Artificial Intelligence, Koblenz, Germany, pp. 320–334 (2005)
16. Bellotto, N., Hu, H.: Vision and Laser Data Fusion for Tracking People with a Mobile Robot. In: International Conference on Robotics and Biomimetics ROBIO 2006, Kunming, China, pp. 7–12 (2006)
17. Siddiqui, M., Medioni, G.: Robust real-time upper body limb detection and tracking. In: Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks VSSN 2006, Santa Barbara, California, USA, pp. 53–60 (2006)
18. Wilhelm, T., Bohme, H.-J., Gross, H.-M.: Sensor Fusion for Visual and Sonar based People Tracking on a Mobile Service Robot. In: Proceedings of the International Workshop on Dynamic Perception 2002, Bochum, Germany, pp. 315–320 (2002)
19. Lastra, A., Pretto, A., Tonello, S., Menegatti, E.: Robust Color-Based Skin Detection for an Interactive Robot. In: Proceedings of the 10th Congress of the Italian Association for Artificial Intelligence (AI*IA 2007), Roma, Italy, pp. 507–518 (2007)
20. Han, J., Bhanu, B.: Fusion of Color and Infrared Video for Dynamic Indoor Environment Human Detection. Pattern Recognition 40(6), 1771–1784 (2007)
21. Muñoz-Salinas, R., Aguirre, E.: People Detection and Tracking Using Stereo Vision and Color. Image Vision Computing 25(6), 995–1007 (2007)
22. Viola, P., Jones, M.J.: Robust Real-time Object Detection, Robust Real-time Object Detection. Cambridge Research Laboratory 24 (2001)
23. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
24. Canny, F.: A computational approach to edge detection. IEEE Trans. Pattern Analysis and Machine Intelligence 8(6), 679–698 (1986)
25. Treptow, A., Cielniak, G., Ducket, T.: Active People Recognition using Thermal and Grey Images on a Mobile Security Robot. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Edmonton, Alberta, Canada (2005)