Detection of Multiple People by a Mobile Robot in Dynamic Indoor Environments

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

National Institute of Astrophysics, Optics and Electronics, Computer Science Department, Luis Enrique Erro 1, 72840 Tonantzintla, México
{polanco,munoz,emorales}@inaoep.mx

Abstract. Detection of multiple people is a key element for social robot design and it is a requirement for effective human-robot interaction. However, it is not an easy task, especially in complex real world scenarios that commonly involve unpredictable motion of people. This paper focuses on detecting multiple people with a mobile robot by fusing information from different sensors over time. The proposed approach applies a segmentation method that uses the distance to the objects to separate possible people from the background and a novel adaptive contour people model to obtain a probability of detecting people. A probabilistic skin model is also applied to the images and both evidences are merged and used over time with a Bayesian scheme to detect people. We present experimental results that demonstrate how the proposed method is able to detect people who is standing, sitting and leaning sideways using a mobile robot in cluttered real world scenarios.

1 Introduction

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 [3]. 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 previous work[2] presents a people detection method which took into account the distance between the robot and people to try to get the best fit model. Therefore, the previous work detected people without assuming that the person is facing the robot. However, the main drawback of that method was the number of false positives detected. In this work, we extend our previous work to: (i) detect multiple people with a mobile robot, (ii) fuse information from different sensors over time, and (iii) apply a Bayesian scheme without a prior visual information about the environment. The sensor fusion significantly reduced the number of false positives in the detection.

A. Kuri-Morales and G.R. Simari (Eds.): IBERAMIA 2010, LNAI 6433, pp. 522–531, 2010.

The paper is organized as follows: In Section 2 we present related work. 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 skin detection method. Section 6 shows the people detection and tracking algorithms. Section 7 shows the experiments and results. Finally, Section 8 presents the conclusions and future research directions.

2 Related Work

Depending on the specific application that integrates people detection and identification there are two principal approaches that can be applied; whole human body detection [4,5] and part-based body detection [6] over images acquired, for instance, with single or stereo charge coupled device (CCD) camera and thermal cameras.

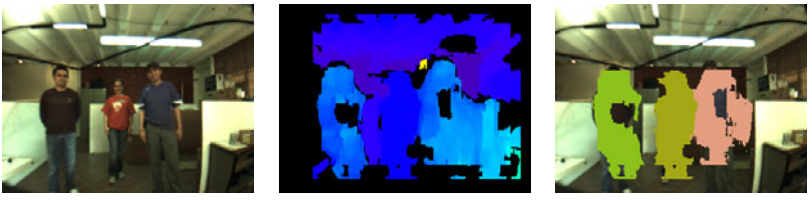
The advantages of whole human body detection approaches are the compact representation of the human body models such as human silhouettes [4,6]. The main drawbacks of these approaches are the sensitivity to object occlusion and cluttered environments as well as the need of robust object segmentation methods. Concerning part-based body detection approaches, the main advantage is their reliability to deal with cluttered environments and object occlusion because of their approach to detect human body parts. In contrast to whole body detection, part-based body detection 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 [7] or skin color to detect hands, arms and faces [8]. Although people detection using static cameras has been a major interest in computer vision [9], some of these works cannot be directly applied to a mobile robot which has to deal with moving people in cluttered environments.

3 Distance Based Segmentation

We use our the proposed method in [2] to segment the information provided by a stereo camera. The distance to the objects can be calculated with an adequate calibration of the stereo camera and using a disparity image. Once the distance to the objects has been calculated, we scan one of the stereo camera images to segment objects based on the obtained depth 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 and Y coordinates). The clusters are defined as follows:

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

where μ_X^k, μ_Y^k and μ_Z^k are the means of the X,Y,Z coordinates of the pixels within the cluster k, σ_X^k, σ_Y^k and σ_Z^k are the variances of the X,Y,Z coordinates



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

Fig. 1. Segmentation method. These examples illustrate how our segment method is able to segment one or multiple people. (a) is the original image, (b) is the disparity image and (c) is the result of our segmentation process.

of the pixels within the cluster k, ρ_k is a vector containing the coordinates of the pixels in the cluster k, and N defines the maximun number of clusters.

After segmentation, clusters with irregular proportions ($\mu_X > \mu_Y$) are eliminated to take into account only objects with human proportions.

Figure 1 shows examples of the object segmentation based on distance using the method described in this section. In this example the background is discarded because it has no enough depth information so, any cluster is created for this region. 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, similar to the model used in [11]. 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:

$$d^t_{body} = (x_T, y_T, w_T, h_T, x_H, y_H, w_H, h_H) \tag{2}$$

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.

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 incomplatible 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 height. To

determine which variation has the best fit, we evaluate the fitting probability as the probability of the parameters of the contour people model given a person p_i, that we denote as $P(d^t_{body}|p_i)$. The probability is calculated as follows:

$$P(d^t_{body}|p_i) = maxValue\frac{L_f(d^t_{body})}{N_T} \tag{3}$$

where L_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 [14].

Finally, since the accuracy of the disparity picture decreases with distance, we reformulated Equation (3) as follows:

$$P(d^v_{body}|p_i) = \begin{cases} \frac{P(d^v_{body}|p_i)}{R\cdot e^{-\rho\mu_z}} & if \;\; \mu_z \leq \delta \\ 0 & if \;\; \mu_z > \delta \end{cases} \tag{4}$$

where R is a normalization parameter, ρ denotes the resolution at which the distance can be measured and δ is the maximum distance allowed. This probability decreases with the distance Z represented by the average μ_z of the distances of all points belonging to region i.

5 Skin Detection

In addition to the elliptical people model, we employed a probabilistic model for skin detection. Skin detection is the process of finding pixels that represent color skin in an image or a video labeling the pixels whether they represent skin or a non-skin pixel. The skin is detected using a color image of the stereo camera and applying the method proposed by Gómez et al. [10]. This method consists on a combination of components of different color spaces. A color space is a model that refers to the color composition in terms of variables such as intensity, luminance, brightness, etc. Some of the most commonly used color models are RGB, HSV, YCrCb, YES, and RGBy. The combination of color spaces obtained by Gomez et al. [10] allows to detect skin in images captured in both indoors and outdoors environments. To select the most appropriate components for detecting skin they used two types of images (i) skin and non-skin in indoor environments, and (ii) skin and non-skin in outdoors environments. The two sets of images covering more than 2000 people of different races and ages. These sets of images were used to obtain 3.350 regions containing skin information. The average size of the regions were 35×35 pixels. By using machine learning tools such as CN2 [13] the following formula was obtained to detect pixels with skin information in an image:

$$SkinDetected = true \; if \; (E > 13.4224) \, and \, ((red/green) < 1.7602) \tag{5}$$

where the component E is calculated as $E = \frac{1}{2}(red - green)$.

The skin detection process using Equation (5) is performed independently in each pixel of the image. This type of detection technique is called pixel-based skin detection, which has the characteristic of not assuming any dependency between the pixels of the image.

5.1 Probabilistic Skin Model

The skin color model determines whether a pixel in the image represents a skin or not. However, it is common to detect false positives in certain regions of the image. Moreover, it is desirable to assign a skin detection probability rather than the true/false output of Equation (5). Therefore, we use a Skin Probability Map (SPM) [12] for skin detection. An SPM is a histogram in which the occurrences of each possible value of the components of a color model are stored. We use two SPMs, one for each set of training images, skin and non-skin. The pixels of each image of the training set were labeled as skin or not using the equation 5. Since the training sets of images were taken from environments similar to those in which the robot is expected to execute their tasks, the color histograms obtained can be considered as probability distribution functions for the detection of skin and non-skin in such environments. This allows not only to obtain a probability for skin detection but also to eliminate false positives in the detection. The RGB color model was used to obtain the histograms for skin $Hist_{skin}$ and non-skin $Hist_{non-skin}$. Thus, if each element of the color histogram is divided by the total number of points it is possible to obtain a probability of skin for a given element [r, g, b]. The conditional probability of a pixel given that it represents skin or non-skin is determined by applying Equations (6) and (7):

$$P(rgb_{(x,y)}|skin) = \frac{Hist_{skin}[r,g,b]}{Total_{skin}} \tag{6}$$

$$P(rgb_{(x,y)}|non-skin) = \frac{Hist_{non-skin}[r,g,b]}{Total_{non-skin}} \tag{7}$$

where $Total_{skin}$ is the total number of elements in the histogram of skin $Hist_{skin}$ and $Total_{non-skin}$ is the total number of elements in the histogram of non-skin $Hist_{non-skin}$.

To obtain the probability of skin given a person p_i, the probabilities of skin on each pixel are averaged according to the image region i to which they belong as follows:

$$P(skin|p_i) = \frac{\sum_{x=0}^{M}\sum_{y=0}^{N} P(rgb_{(x,y)}|skin)}{N \times M} \tag{8}$$

where $M \times N$ is the size of the region containing the person p_i, and $rgb_{(x,y)}$ is the RGB value for the pixel (x, y).

To calculate $P(skin|p_i)$ we only consider the pixels that were detected with the SPM as skin using Equation (9):

$$\frac{P(rgb_{(x,y)}|skin)}{P(rgb_{(x,y)}|non-skin)} \geq \theta \tag{9}$$

where θ is obtained experimentally as 0.8 and represents the threshold at which can be determined the presence of skin.

In a similar way to the elliptical model, the distance at which the region is detected is considered in the skin model since the accuracy of the disparity image decreases with distance. Therefore, the final expression for calculating the probability of people detection based on our skin model is defined as follows:

$$P(skin|p_i) = \begin{cases} \frac{P(skin|p_i)}{R \cdot e^{-\rho\mu_z}} & if \;\; \mu_z \leq \delta \\ 0 & if \;\; \mu_z > \delta \end{cases} \tag{10}$$

where R is a normalization parameter, ρ denotes the resolution at which the distance can be measure and δ is the maximum distance allowed. This probability decreases with the distance Z represented by the average μ_z of the distances of all points belonging to region i.

6 People Detection and Tracking

At this point, we can determine with some probability, the presence of people in an image using the contour of people and skin evidence independently. The integration of the evidence was performed by a weighted sum of their probabilities using Equation (11):

$$P(skin, contour|p_i) = \alpha \times P(skin|p_i) + (1-\alpha) \times P(d^v_{body}|p_i) \tag{11}$$

where α is a weight associated with evidence of skin and silhouette which is experimentally determined as 0.4.

In order to improve the detection process we use video streaming and combine evidence of several frames with a Bayesian scheme. Once a person has been detected at time t , we proceed to search if that person was previously detected, with our proposed method, at time $t-1$ calculating the Euclidean distance from current regions that contains a person to previous regions. If the distance between two regions is less than a threshold value, then we consider these people to be the same. To calculate the probability of a person p_i at time t we apply

$$P(p_i^t) = \frac{P(skin, contour|p_i^t)P(p_i^{t-1})}{P(skin, contour|p_i^t)P(p_i^{t-1}) + P(skin, contour| \sim p_i^t)P(\sim p_i^{t-1})} \tag{12}$$

where $P(skin, contour|p_i^t)$ is the probability of skin and contour given the person p_i at time t which is calculated applying equation 11. $P(p_i^{t-1})$ is the probability of the person p_i at time $t-1$. $P(skin, contour| \sim p_i^t)$ is the probability of skin and contour given that the person p_i was not detected at time t and $P(\sim p_i^{t-1})$ is the probability of that the person p_i has not been detected at time $t-1$. In the experiments the values of $P(p_i^t)$ when $t=0$ were set to 0.5.

7 Experimental Results

We present experimental results that demonstrate how the proposed method, using an ActivMedia PeopleBot mobile robot equipped with a stereo camera, is able to detect people who is standing, sitting and leaning sideways at 12 f.p.s. using both real-time and pre-recorded video. Moreover, it can also detect people from their backs and not facing the robot. In order to test and evaluate our method we use a mobile robot in dynamic indoor environments with different lighting conditions and without assuming a prior visual model of the environment. People was placed at different distances from the robot (1 to 5 m). The number of people varies from one to three.

Figure 2 shows how our people detection method is able to detect multiple people and track them over time. The experiments were performed with a maximum of three people; however, in Figure 2 we show an experiment with two people to compare the performance of the people detection method using a single frame detection scheme against our people detection method using evidence from several frames applying a Bayesian scheme. Snapshots of the experiments with three people are shown in Figure 4. We calculated the detection rate DR as follows:

$$DR = \frac{N_D}{N_T} \tag{13}$$

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 considered.

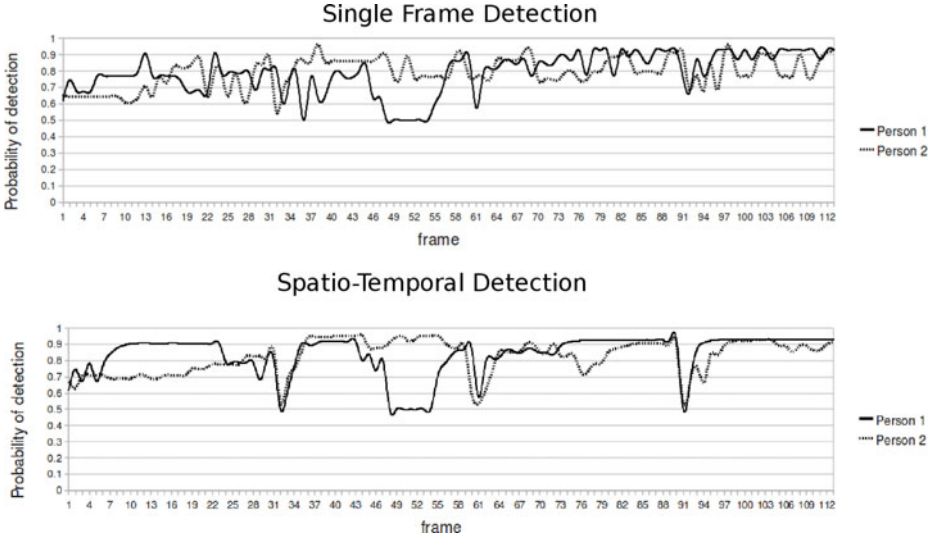


Fig. 2. Multiple people detection performance using single and spatio-termporal detection

Our segmentation method has a DR of 94.0% with one false positive every 1800 regions examined.

Table 1 presents the detection rates of our method merging the evidence of silhouette and skin. A closely related work is proposed in [11], but since they use a thermal camera for people detection, a direct comparison is not possible (their detection rates presented in [11] are in the order of 91.9). Therefore Table 1 only shows a comparison with the results presented in [2]. The detection rate using both skin and contour evidence is lower than the detection rate obtained by using only contour evidence. However, as can be seen in Figure 3, by using only the contour people model we can obtain false positives on the background (Figure 3 (a) and (c)) whereas with our proposed method, fusing the evidences of contour and skin, those false-positives can be successfully filtered.

In Figure 4 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 which is also moving in real world scenarios using our proposed people detection method.

Table 1. Detection Rates and False Positives

Method	Detection rate	False Positives / regions examined
Adaptive countour model [2]	89.0	53 / 1800
Adaptive spatial-temporal model with contour model [2]	96.0	32 / 1800
Our method (Merging contour and skin)	94.0	1 / 1800

(a) (b) (c) (d)

Fig. 3. False positives reduction

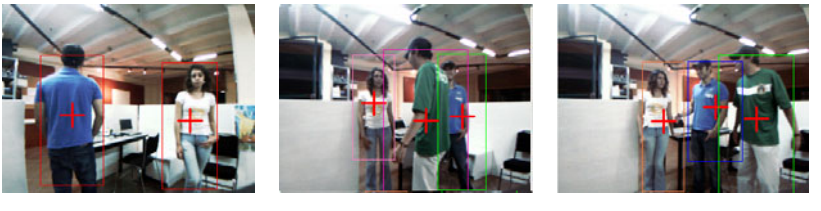


Fig. 4. Experiments performed using a mobile robot in dynamic indoor environments

8 Conclusion and Future Work

This paper focuses on detecting multiple people with a mobile robot by fusing information from different sensors over time and applying a Bayesian scheme without a prior visual information about the environment. The proposed segmentation method takes into account the distance between the robot and people to perform a segmentation and to provide a first estimation of people location. We use an adaptive contour people model based on people distance and a color model of skin to calculate a probabiliy of people etection, one for each evidence. To detect people, we merged the probabilities of the fused model over time by applying a Bayesian scheme. The experimental results show how our proposed method is able to segment and detect standing and sitting people, as well as people in frontal and side view with a mobile robot without assuming a previous visual model of the environment or that people are facing or not the robot. By using only the contour people model we can obtain false positives on the background, whereas, with our proposed method, fusing the evidences of contour and skin, those false-positives can be successfully filtered. As future research work we are considering not only detect people but also recognize them.

References

1. Burgard, W., Cremers, A., Fox, D., Hhnel, D., Lakemeyer, G., Schulz, D., Steiner, W., Thrun, S.: Experiences with an Interactive Museum Tour-guide Robot. Artificial Intelligence, 3–55 (1999)
2. Méndez-Polanco, J.A., Muñoz-Meléndez, A., Morales, E.F.: People Detection by a Mobile Robot Using Stereo Vision in Dynamic Indoor Environments. In: Gelbukh, A., de Albornoz, Á., Terashima-Marín, H. (eds.) MICAI 2005. LNCS (LNAI), vol. 3789, p. 349. Springer, Heidelberg (2005)
3. Osada, J., Ohnaka, S., Sato, M.: The Scenario and Design Process of Childcare Robot, PaPeRo, Dynamic Indoor Environments, pp. 80–86 (2006)
4. Malagón-Borja, L., Fuentes, O.: Object Detection Using Image Reconstruction with PCA. Image Vision Computing 27(1-2), 2–9 (2009)
5. 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)
6. Müller, S., Schaffernicht, E., Scheidig, A., Hans-Joachim, B., Gross-Horst, M.: Are You Still Following Me? In: Proccedings of the 3rd European Conference on Mobile Robots ECMR, Germany, pp. 211–216 (2007)
7. Schaffernicht, E., Martin, C., Scheidig, A., Gross, H.-M.: A Probabilistic Multimodal Sensor Aggregation Scheme Applied for a Mobile Robot. In: Proccedings of the 28th German Conference on Artificial Intelligence, Koblenz, Germany, pp. 320–334 (2005)
8. 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)
9. Han, J., Bhanu, B.: Fusion of Color and Infrared Video for Dynamic Indoor Environmentsving Human Detection. Pattern Recognition 40(6), 1771–1784 (2007)

10. Gómez, G., Sánchez, M., Sucar, L.E.: On Selecting an Appropriate Colour Space for Skin Detection. In: Coello Coello, C.A., de Albornoz, Á., Sucar, L.E., Battistutti, O.C. (eds.) MICAI 2002. LNCS (LNAI), vol. 2313, pp. 69–78. Springer, Heidelberg (2002)
11. 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)
12. Jones Michael, J., Rehg James, M.: Statistical Color Models with Application to Skin Detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 1999), vol. 1, p. 1274–1280 (1999)
13. Clark, P., Boswell, R.: Rule induction with CN2. In: Kodratoff, Y. (ed.) EWSL 1991. LNCS (LNAI), vol. 482, pp. 151–163. Springer, Heidelberg (1991)
14. Canny, J.: A Computational Approach To Edge Detection. IEEE Transactions Pattern Analysis and Machine Intelligence (1986)