# Person Following through Appearance Models and Stereo Vision Using a Mobile Robot

Daniele Calisi, Luca Iocchi and Riccardo Leone

Dipartimento di Informatica e Sistemistica
University ”La Sapienza”, Roma, Italy

**Abstract.** Following a person is an important task for mobile service and domestic robots in applications in which human-robot interaction is a primary requirement. In this paper we present an approach that integrates appearance models and stereo vision for efficient people tracking in domestic environments. Stereo vision helps in obtaining a very good segmentation of the scene to detect a person during the automatic model acquisition phase, and to determine the position of the target person in the environment. A navigation module and a high level *person following* behavior are responsible for performing the task in dynamic and cluttered environments. Experimental results are provided to demonstrate the effectiveness of the proposed approach.

## 1 Introduction

Following a person is an important task for mobile service and domestic robots in applications in which human-robot interaction is a primary requirement. Consequently, many works have been presented in the literature for people tracking and following from a mobile platform. Most of these works use vision as the main sensor for detecting, tracking and localizing people in the environment. Face detection (e.g., [8, 4]), skin color detection (e.g., [8, 11]), and color histograms (e.g., [12]) are the most common methods that have been considered. Stereo vision has also been used to better determine the location of the person with respect to the robot. In particular, stereo vision has been used in [5] to build a floor plan occupancy map as background model, that is used to determine moving objects in the scene, while tracking is obtained by Kalman Filtering; independent motion estimation from disparity images is computed in [1] to detect a moving target from a moving platform; and integration of stereo matching with face detection and skin color detection is presented in [8]. Besides vision, some approaches using measures from laser range finders to detect people legs have been also proposed [14, 15] as well as integration of LRF and vision [4]. In addition, the works in [14, 15] are mostly focussed to the problem of tracking multiple people, presenting approaches based on Particle Filters. Finally, active sensors (such as RFID, or light-emitting devices) can be used for this task, but this requires the person to wear such devices and it may be not suitable in some applications.

Regardless the amount of work in this field, a complete reliable solution that can be used in general conditions is not currently available. Face detection and skin color

Calisi D., Iocchi L. and Leone R. (2007).
Person Following through Appearance Models and Stereo Vision Using a Mobile Robot.
In *Robot Vision*, pages 46-56
DOI: 10.5220/0002069800460056

detection are only available when the person is oriented towards the robot, and this is not common for a person following task. Approaches based on color histograms often require a priori knowledge (or somewhat manual pre-calibration) about the person and are sensitive to other colors present in the environment. Moreover, the use of a single camera makes it difficult to evaluate the location of a person in the environment, especially in presence of other objects between the robot and the person. Using range information provided by stereo overcomes some of the previous difficulties, but it may have problems with cluttered environments. Finally, approaches using LRF to detect people legs can be easily confused by other elements in the environment (such as tables and chairs) and only if the robot is correctly localized in a known map and such objects are static, these false observations can be filtered out.

In this paper we focus on a reference scenario, that is quite challenging for many of the works presented above. A person must be followed by the robot in a domestic environment. We assume the robot to have a map of the environment, so to localize itself, but we also consider the presence of other people moving around and objects not represented in the map that can also stay between the robot and the person. No a priori knowledge about the person to be followed is available, and the person will move in a natural way in the environment (this implies for example that his/her face will be rarely visible). Finally, we assume that the task is started with nobody in the field of view of the sensors of the robot (in particular, of the camera) and that the first person entering the scene is the one that the robot will have to follow. This scenario is partially borrowed from the RoboCup@Home competitions[1].

The contribution of this paper is presenting an integrated approach for person modelling, tracking and following that is based on building appearance models of the person to be tracked, using them for detecting the target person, and then using stereo vision computation for 3D modelling of the scene and person location. The advantages in using such an integrated approach are in the combination of two important features: person recognition and tracking based on appearance models, 3D localization and tracking of the person based on stereo vision. In fact, appearance models can reliably detect and track people in the image space, but have problems in determining their position in the environment for driving a robot navigation process. On the other hand, stereo vision based people tracking, as well as other 2D geometric-only approaches, present some limits when multiple people are close in the environment or when objects in the environments (e.g., chairs, tables) have geometric properties similar to people.

The approach presented in this paper uses appearance models to filter perception, in order to detect in the image space those pixels that belong to the target person and then uses stereo vision analysis for determining ground position of the person and to drive the robot navigation process. The integrated approach is therefore suitable in presence of multiple people and in cluttered environments. More specifically, the proposed approach works in two phases: i) in the first phase we assume the robot is still and only the target person is in the field of view of the stereo camera, during this time the model of the person is automatically acquired; ii) in the second phase the acquired model is used for image segmentation and stereo computation is used to determine the location of the target person in the environment, while the robot follows him/her.

[1] `http://www.robocupathome.org/`

Notice that, in the scenario presented above, it is not really necessary to track multiple people, but it is sufficient to track one target person to be followed, while considering other people as obstacles to be avoided during navigation. Although it is in principle possible to apply one of the methods for tracking multiple people, it seems more convenient to implement a specific solution that do not solve explicitly a multi-target tracking problem, but just a multi-hypothesis single person tracking problem.

One major drawback of the system is that it is not able to distinguish people wearing clothes similar to the target person or it may be confused in presence of other objects of similar colors of the target person. Although a filter considering the global position of the observations allows for filtering out observations that are too close or behind the walls in the map, there may still be cases in which false positives affect the performance of the robot. Besides this issue, that is common to color based recognition approaches, our method works reasonably well in the reference scenario described above.

We have implemented the described approach on top of a Pioneer robot equipped with a stereo vision camera and a laser range finder. The laser range finder is used only for localization and obstacle detection, while the stereo camera is used only for people modelling, detection and tracking. Experimental results both in a simulated environment and in a real domestic-like environment provide evidence of the effectiveness of the proposed method.

The paper is organized as follows. In Section 2 we describe appearance model acquisition, and in Section 3 person detection and tracking. Implementation details and experiments are described in Section 4, and conclusions are drawn in Section 5.

## 2 Image Segmentation and Appearance Model Acquisition

As already mentioned, the method presented in this paper is executed in two phases. The first phase consists in the acquisition of an appearance model of the person that will be followed by the robotic platform and it is executed with the robot still and under the assumption that only the target person will appear in front of the robot.

Under these assumptions, this process corresponds to people localization and tracking process with a fixed stereo camera (PLT), as described in [3, 10]. The PLT system has been thus used for the first phase, providing for segmenting the image by distinguishing the foreground person from the background, by background subtraction considering both intensity and disparity (see [3] for details).

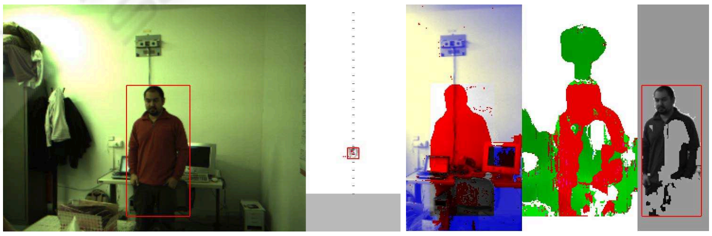


**Fig. 1.** Examples of segmentation provided by the stereo tracker.

Figure 1 shows an example of such a process: on the left side the person detected in the scene and his location projected on the plan-view, on the right side the two components of intensity and disparity segmentation, and the extracted foreground.

More specifically, for each tracked person in the scene, the PLT system provides a set of data $\Omega = \{\omega_{t_0}, ..., \omega_t\}$ from the time $t_0$ in which the person is first detected to current time $t$. The value $\omega_t = \{(X_t^i, Y_t^i, Z_t^i, R_t^i, G_t^i, B_t^i) | i \in \mathcal{P}\}$ is the set of XYZ-RGB data (i.e. 3D location and color) for each pixel $i$ that is classified as belonging to a person $\mathcal{P}$. The reference system is chosen in order to have the plane XY coincident with the ground floor.

From the segmentation provided by the PLT module, we can build an appearance model for the person to be tracked. Many existing appearance models have been proposed in the literature. For example, [9] propose a Texture Temporal Template in which the color information over time for every pixel of the silhouette is used to build the person model. A probabilistic approach is described in [16], while in [7] the model is extended in order to be robust to changing lighting conditions.

Among all these models, we choose to model the appearance of a person with two uni-dimensional color distributions. Our choice has been mainly motivated by efficiency reasons, since during the second phase of our method the appearance model must be used as a filter for the image pixels at every frame and thus such a filter must be implemented in real-time. Moreover, 3D information about the pixel location is important to filter out pixels whose height from the ground is below 1 meter, thus focussing only on the upper part (torso) of the person. This decreases probability of integrating in the model false readings (i.e., colors not belonging to the person). Other models can be taken in consideration, as long as a real-time filter is available for the second phase.

More specifically, the appearance model is composed by two distributions $\mathcal{A} = \langle D_H(), D_V() \rangle$, where $D_H(h)$ and $D_V(v)$ are uni-dimensional color distributions respectively in the H component and in the V component of the HSV color space, and are defined by

$$D_H(h) = \frac{\sum_{\tau=t_0}^{t} |\omega_\tau^{(h)}|}{\sum_{\tau=t_0}^{t} |\omega_\tau|} \qquad D_V(v) = \frac{\sum_{\tau=t_0}^{t} |\omega_\tau^{(v)}|}{\sum_{\tau=t_0}^{t} |\omega_\tau|}$$

where $\omega_\tau^{(h)}$ ($\omega_\tau^{(v)}$) is the subset of $\omega_\tau$ containing only those pixels whose component $H = h$ (respectively, $V = v$).

Examples of color distributions for three people used in the experiments are given in Figure 2. When the target person wears a uniform color, the above distributions are very peaked and the values in the color distributions can also be seen as the probability that a pixel of a given color belongs to the target person. In case of multiple colors in the person appearance, the above color distributions do not represent anymore probability distributions over the color space of pixels belonging to the target person. If a person wears $k$ principal colors, then $k$ color distributions should be built one for each principal color, by appropriately clustering these data, and then a logical disjunction among these distributions is applied during the filtering phase.

In our approach we use a representation based on a single color distribution, thus assuming that people wears uniform colors. This is very common in practice and works

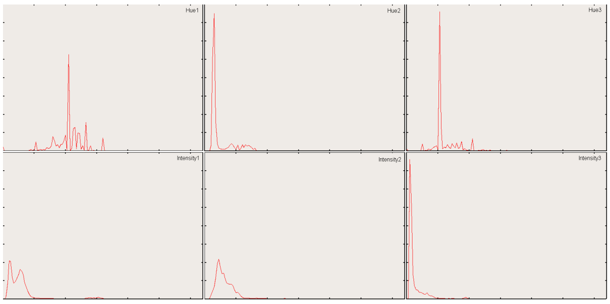


**Fig. 2.** Appearance models from three people.

well also because we consider only pixels that are above 1 m from the ground. When this assumption is not verified, i.e. in case of people wearing clothes that do not have a principal color (e.g., multi-colored shirts), the color distribution has multi modes and tend to be less peaky, therefore it is more likely to have false positives in the filtering phase.

Moreover, observe that hue is not defined when saturation is zero and generally not relevant for low saturated colors, as well as value is not relevant for well saturated colors. This is taken into consideration in the filtering phase (see next section), by weigthing the use of the two distributions with the saturation of each color pixel.

It is also interesting to observe that an important feature of the PLT system is to provide a very good segmentation of the image, since the combination of intensity and disparity allows for coping with typical problems of background subtraction techniques, such as shadows, reflections, ghosts, etc. Consequently, the contamination of the appearance model given by pixels that do not belong to the person is very limited.

After a predefined time the appearance model $\mathcal{A}$ of the target person is stored in memory and the second phase can start. Usually, only a few seconds are needed for the first phase. In our experiments, we used 5 seconds with empty scene for fast background modelling, and 10 seconds with the person in the field of view of the camera for appearance model acquisition. Since the PLT system runs at about 10 frames per second, we use about 50 frames for background modelling and 100 frames for person modelling.

## 3 Person Recognition and Tracking

During the second phase of our process the robot exploits the appearance model of the target person to detect him/her in the scene, to compute its relative location, and to follow him/her. The second phase is composed by four steps: 1) the appearance model is used as a filter on the current image to detect those pixels that are consistent with the model (i.e., candidate to belong to the target person); 2) stereo computation is applied on these pixels and a plan-view is generated from 3D data about these pixels; 3) plan view is processed to detect connected components (*world blobs*); 4) world blobs are tracked over time and periodically sent to the navigation module.

It is interesting to observe here that the steps 2 to 4 are again taken by the PLT system [3, 10]; in fact, only the segmentation process of the original PLT system must be replaced when cameras are mounted on a mobile platform, since it is not possible anymore to rely it on a background model. Therefore we replace in PLT the segmentation based on background subtraction with a filtering based on the appearance model determined in the first phase.

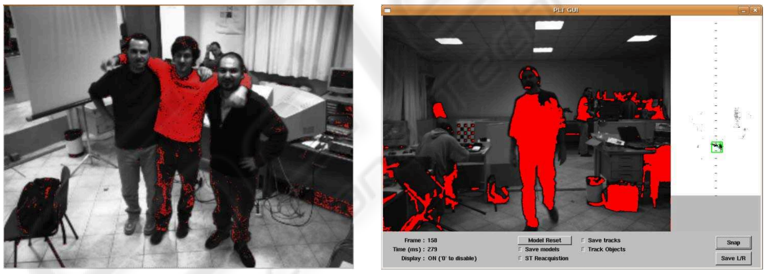
### 3.1 Image Filtering Based on Appearance Model

The appearance model $\mathcal{A} = \langle D_H(), D_V() \rangle$ is used to determine a set of foreground pixels $\mathcal{F}$, as follows

$$\mathcal{F} = \{ i \in \mathcal{I} | s(i) D_H(h(i)) + (1 - s(i)) D_V(v(i)) > \delta \}$$

where for each pixel $i$ in the current image $\mathcal{I}$, we select it as a foreground pixel if and only if the value $s(i)D_H(h(i)) + (1 - s(i)D_V(v(i))$ is above a given threshold $\delta$. In the above expression $h(i), s(i), v(i)$ are the values of the HSV components of the pixel $i$ and the value of $s(i)$ is used as a weight factor between the two distributions $D_H(h(i))$ and $D_V(v(i))$. This is motivated by the fact that for high saturated colors the contribution of the H component is more relevant, while for low saturated ones, the V component is more significant.

Notice that this is a pixel-based operation that can be implemented in a very efficient way using look-up tables.



Frame : 158
Time (ms) : 279
Display : ON ('0' to disable)

**Fig. 3.** Examples of color based filtering.

Examples of image segmentation based on color filtering is shown in Figure 3. The threshold $\delta$ can be set either manually from empirical tests or automatically with an adaptive procedure. In this second case, it is sufficient to determine, at the end of the first phase described in the previous section, the value of $\delta$ that minimizes false positives and false negatives, where false positives are counted as the number of pixels that will be detected as foreground and do not belong to the person and false negatives are those pixels that belong to the person but whose expression is not above the threshold. In our implementation, at the end of the first phase we determine the threshold $\delta$ by starting with a low value and incrementing it until the number of false positives increase above a given amount from its initial value.

### 3.2 Plan View Analysis and Tracking

After image segmentation, a plan view analysis is performed as described in [3]. This step maps foreground points to a plan view and detects *world blobs* by connected component analysis, in order to remove segmentation noise due to similar colors. An example of such analysis is given in Figure 3 right side. The person being tracked has very dark clothes so many pixels not belonging to the person are detected by the color based image filter. However, plan-view analysis allows for detecting only one world blob corresponding to the person (as shown in the rightmost diagram).

This mechanism substantially increases the capability of the system to deal with noise due to the presence in the environment of colors similar to those belonging to target person. Moreover, blobs in plan view are also analyzed to assign a score to the observation. This score is defined as the sum of *weights* of the cells belonging to each world blob, and the weight for each cell in the plan view is given by the maximum height (i.e., $Z$) of pixels projected in the cell. World blobs with higher score are more likely to come from the target person, while world blobs with lower score are likely to be false positives.

World blobs are tracked using a multiple target tracker that is implemented with a set of Kalman Filters. Data association is based on minimum distances between observations and tracks, but it also takes into account geometric filtering and world blob scores. The former is used to filter observations that are too close or behind the walls in the environment, allowing for avoiding false positives due to parts of the environment with similar color than the target person. The latter is used by preferring high score targets in presence of many of them.

## 4 Implementation and Experiments

The system described in this paper has been implemented and experimented on a mobile robot acting in the simulated domestic environment. This section presents implementation details and some experimental results.

*Self-Localization.* Since the map of the environment is known, laser range finder based Monte Carlo Localization has been implemented (details can be found in [2]).

*Navigation and obstacle avoindance.* The navigation subsystem follows a two-level approach: the higher lever is a global path-planner [6] that uses as input the whole map to find a feasible global path, the lower level is an obstacle avoidance module that takes as input the current sensor readings and steers the robot towards the goal, avoiding the obstacles (similar to [13]).

*Behaviour specification.* The *Person Following* behavior has been described by a Petri Net Plan (PNP), which is is a formalism for describing high-level activities of agents successfully applied on mobile robots [17]. In this plan the person is followed until he/she is detected by the system. When the robot is within 1 meter from the person, it stops while keeping on tracking him/her (possibly rotating on itself). If the person gets lost, the robot keeps on going to the last position where the person was seen and then begins a seek action until it finds the tracked person again. This seek action simply consists in turning on place towards the direction where the person has been seen the last time. The task fails if after a complete 360 degrees rotation the robot cannot find the person (thus no search is implemented here).

Experimental results that are divided in three groups. First, we have evaluated the vision module and in particular the ability of correctly modelling different people and of recognizing a target person among many people. Second, we have evaluated the *person following* behaviour in a simulated environment[2], with a simplified vision process. Finally, we have experimented the entire system on the mobile robot.

### 4.1 Experiments on the Vision Module

The first set of experiments have been performed in order to estimate the validity of the color based appearance model that has been used in our system. In this experiment only the vision part of the system was active: i.e., appearance model acquisition and person detection. To this end, we have built 5 models from 5 different people in our lab, and then recorded 5 other videos from the same people. We have run the system $5 \times 5$ times, considering all the possible combinations of videos of people and models.

The result of this preliminary test was that all the people have been correctly detected by their own models, thus no false negatives have been registered. More specifically, the system was able to correctly detect each person in the video using the model previously acquired for the same person. Among the other 20 runs, the number of false positives has been 7. These corresponds to situations in which a person was detected by the system using the model of another person. This high rate of false positives was mainly due to the presence of two people having similar clothes. If we remove one of them from this set we obtain only 3 false positives.

For the incorrect cases, we have extended the experiment recording a video with two people: the one the was incorrectly determined and the one corresponding to the used model. In all the cases the target person got the best score, with respect to the other one. Therefore, the robot is able to correctly follow the target person, even in presence of other people with similar colors, as long as the target person remains in the field of view of the stereo sensor.

### 4.2 Experiments of the Entire System on the Robot

The experiments on the real robot confirm the effectiveness of the developed system. In this section we report the results of three experiments, in the environment shown in

[2] not reported here for lack of space.

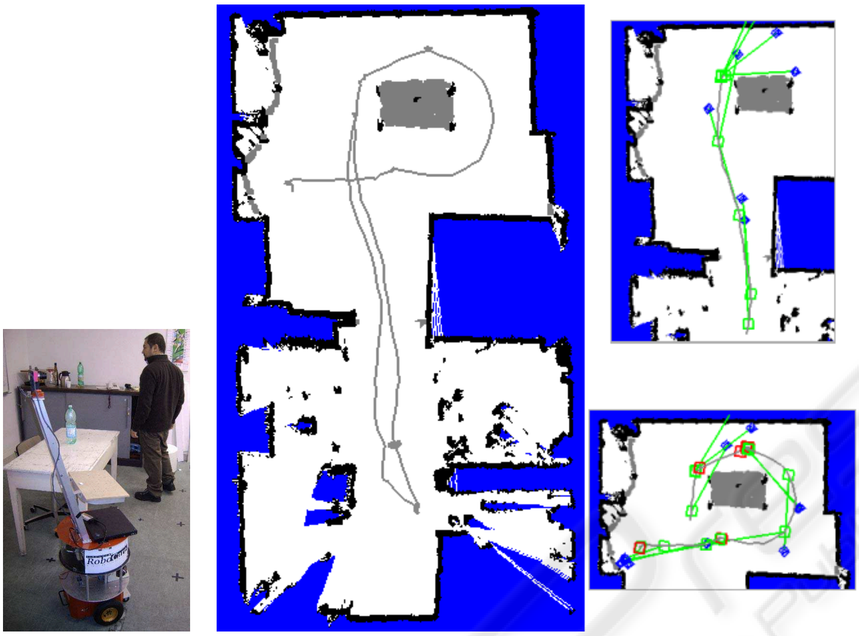


**Fig. 4.** The real environment used in the experiment and the path of the robot and some snapshots.

Figure 4. During experiments 2 and 3 a second person was moving in the environment and the robot was successfully able to avoid him. No contacts between the robot and people or objects in the environment have been registered during the experiments.

In Figure 4, we show a snapshot of the environment and of the robot, the entire map with the path covered by the robot during one of the experiments (Experiment 2 in Table 1), and some details of the person following behaviour (little squares are the robot positions, while the connected diamonds are the person's position detected in that time slot).

**Table 1.** Tests in real scenario.

| Experiment | Distance covered | Time elapsed | Seek time | Avg. speed<br>(when not seeking) |
|---|---|---|---|---|
| Experiment 1 | 24.38 m | 290.41 s | 165.35 s | 0.19 m/s |
| Experiment 2 | 18.17 m | 212.98 s | 131.21 s | 0.22 m/s |
| Experiment 3 | 20.02 m | 206.20 s | 122.81 s | 0.24 m/s |

In Table 1 we show the results of the experiments in the real scenario, with total covered distance and total execution time. In the experiments the maximum robot speed was limited to 0.3 m/s, which is the normal operation speed for this kind of robot in such an environment. The second and third columns show respectively the total length of the path and the total execution time. The fourth column shows the time elapsed during seek actions (so when the robot is not moving), and the fifth column average speed of the robot while it was moving (excluded seek time).

The performed exeriments have confirmed the effectiveness of the method in such an environment.

## 5 Conclusions and Future Work

In this paper we have presented an approach to person following from a mobile robot equipped with a stereo camera, using automatic apperance model acquisition and stereo vision based tracking. The implemented system works well in environments with unknown objects and other moving people, and do not require a priori calibration on the person to be tracked, using instead an automatic fast training step. The main limitation of the approach is given by the color based model, that fails to correctly detect a person when there are other objects or people using similar colors. As future work, we intend to perform more extensive experiments to compare and evaluate different color based appearance models that can be used for this task.

## References

1. M. Agrawal, K. Konolige, and L. Iocchi. Real-time detection of independent motion using stereo. In *Proc. of IEEE Workshop on Motion*, 2005.
2. S. Bahadori, A. Censi, A. Farinelli, G. Grisetti, L. Iocchi, D. Nardi, and G. D. Tipaldi. Particle based approaches for mobile robot navigation. Proc. of the second RoboCare Workshop, Roma, Italy, 2005.
3. S. Bahadori, L. Iocchi, G. R. Leone, D. Nardi, and L. Scozzafava. Real-time people localization and tracking through fixed stereo vision. In *Proc. of Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE)*, pages 44–54, 2005.
4. N. Bellotto and H. Hu. Multisensor integration for human-robot interaction. *The IEEE Journal of Intelligent Cybernetic Systems*, 1, July 2005.
5. D. Beymer and K. Konolige. Real-time tracking of multiple people using stereo. In *Proc. of IEEE Frame Rate Workshop*, 1999.
6. D. Calisi, A. Farinelli, L. Iocchi, and D. Nardi. Autonomous navigation and exploration in a rescue environment. In *Proc. of the 2nd European Conference on Mobile Robotics (ECMR)*, pages 110–115, Edizioni Simple s.r.l., Macerata, Italy, September 2005.
7. R. Cucchiara, C. Grana, G. Tardini, and R. Vezzani. Probabilistic people tracking for occlusion handling. In *Proc. of 17th Int. Conf. on Pattern Recognition (ICPR'04)*, 2004.
8. T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. *International Journal of Computer Vision*, 37(2):175–185, 2000.
9. I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 2000.
10. L. Iocchi and R. C. Bolles. Integrating plan-view tracking and color-based person models for multiple people tracking. In *Proc. of IEEE International Conference on Image Processing (ICIP'05)*, 2005.

11. M. Kleinehagenbrock, S. Lang, J. Fritsch, F. Lömker, G.A. Fink, and G. Sagerer. Person tracking with a mobile robot based on multi-modal anchoring. In *Proc. of IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN)*, 2002.
12. H. Kwon, Y. Yoon, J. B. Park, and A. C. Kak. Person tracking with a mobile robot using two uncalibrated independently moving cameras. In *Proc. of of IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2005.
13. J. Minguez and L. Montano. Nearness diagram (nd) navigation: Collision avoidance in troublesome scenarios. *IEEE Transactions on Robotics and Automations*, 20(1):45–59, 2004.
14. M. Montemerlo, S. Thrun, and W. Whittaker. Conditional particle filters for simultaneous mobile robot localization and people-tracking. In *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2002.
15. D. Schulz, W. Burgard, D. Fox, and A. B. Cremers. People tracking with a mobile robot using sample-based joint probabilistic data association filters. *International Journal of Robotics Research (IJRR)*, 2003.
16. A. W. Senior. Tracking with probabilistic appearance models. In *Proc, of ECCV workshop on Performance Evaluation of Tracking and Surveillance Systems (PETS)*, pages 48–55, 2002.
17. V. A. Ziparo and L. Iocchi. Petri net plans. In *Proc. of Fourth International Workshop on Modelling of Objects, Components, and Agents (MOCA)*, Turku, Finland, 2006.