Cooperative Passers-by Tracking with a Mobile Robot and External Cameras

A. A. Mekonnen[a,b], F. Lerasle[a,b], A. Herbulot[a,b]

[a] *CNRS, LAAS, 7, Avenue du Colonel Roche, F-31400 Toulouse, France*
[b] *Univ de Toulouse, UPS, LAAS, F-31400 Toulouse, France*

Abstract

In this paper, we present a cooperative passers-by tracking system between fixed view wall mounted cameras and a mobile robot. The proposed system fuses visual detections from wall mounted cameras and detections from a mobile robot–in a centralized manner–employing a "tracking-by-detection" approach within a Particle Filtering strategy. This tracking information is then used to endow the robot with passers-by avoidance ability to facilitate its navigation in crowds during the execution of a person following mission. The multi-person tracker's ability to track passers-by near the robot distinctively is demonstrated through qualitative and quantitative off-line experiments. Finally, the designed perceptual modalities are deployed on our robotic platform, controlling its actuators via visual servoing techniques and free space diagrams in the vicinity of the robot, to illustrate the robot's ability to follow a given target person in human crowded areas.

Keywords: Multi-target Tracking, Multi-sensor Fusion, Automated Person Detection, Cooperative Perception Systems

1. Introduction

Currently, there is an increasing demand for the deployment of service robots in public all day human environments. In an effort to fuel this demand, various researchers have deployed prototype robotic systems in populated environments like hospitals [1], supermarkets [2], museums [3], and

Email addresses: aamekonn@laas.fr (A. A. Mekonnen), lerasle@laas.fr (F. Lerasle), aherbulo@laas.fr (A. Herbulot)

office environments [4] to name a few. A core issue that should not be overlooked, as robots come out of their isolated industrial milieu and start interacting with humans in a shared workspace, is safe interaction. For a service robot to exhibit safe and acceptable interactive behavior–it needs to know the presence, whereabouts, and movements of people to better understand and anticipate their intentions and actions. In general, a service robot should have two main capabilities (i) identifying (discriminating) interesting person(s) to interact with–depending on the task at hand, and (ii) carrying out the task with as minimal interference as possible with the other persons in the workspace. The first capability necessitates an active interaction task that requires the robot to clearly identify a subject for interaction and engage with him/her. The interaction could, for example, be following or intercepting that person. The second capability entails a passive interaction with the passers-by in the environment. The robot should consider them as dynamic obstacles with special requirements and control its maneuver so as to avoid them. To achieve all this, the robot needs to have a robust human perception and motion control functionality.

Though extremely desirable, this coupled human perception and motion control task is a very complex and challenging task to accomplish from a mobile platform. Automated person detection is by far one of the most challenging tasks due to physical variation of persons, body deformations, appearance variation due to different clothings, etc. These challenges are further amplified in mobile robots because of mobility of the sensors, limited Field-Of-View (FOV) of on-board sensors, and limited on-board computational resources. These challenges are partly alleviated by fusing various sensor outputs to make best informed decisions, *e.g.,* [5, 6, 7]. The extent of the improvement depends on how well the different sensors complement each other. Generally, the more complementary information the different sensors provide, the better the perception [8]. Recently, some researchers have considered cooperative perception systems using a mobile robot and wall mounted camera(s) [9, 10]. This opens up more possibilities as it allows the system to benefit from the advantages of both perception modes: perception from wall mounted cameras and perception from the mobile platform. For instance, the system will have global perception from the wall-mounted cameras which lead to increased anticipation capabilities; on the other hand, the mobile platform provides local perception, a means for action, and (as it can move around) the ability to cover dead spots and possibly alleviate occlusions. These kinds of cooperative systems have the potential to lead to

more generic surveillance systems as they can handle various scenarios.

In this paper, we present a cooperative perception system made up of wall mounted cameras and a mobile robot to perceive passers-by in a surveilled area. The purpose is not motion capture and/or articulated motion analysis, rather, it is to obtain the positions and trajectories of passers-by on the ground plane. This work builds upon our previous work on a person following mobile robot presented in Germa *et al.* [5]. Our goal is to realize a person following mobile robot that will follow a single target (user) person wearing an RFID tag while at the same time taking the dynamics of the passers-by into consideration to adjust its motion accordingly to avoid them. In [5], the functionalities to detect/track a target person (based on vision and RFID) and follow him/her robustly have been exhaustively addressed. This work addresses and thoroughly presents the missing functionalities: passers-by perception via a cooperative perceptual system and associated control law for avoidance. The remainder of this paper is structured as follows: Section 2 presents general overview and related works. Section 3 presents framework and architecture of the overall envisaged system in detail. Consequently, section 4 explains the different multi-person detection modalities that drive the passers-by tracking (presented in section 5). Sections 6 and 7 present offline evaluation of the passers-by tracking in various modes; and passers-by avoidance, integration details of the developed functionalities on the target robotic platform, and associated live runs, respectively. Finally, the paper concludes with concluding remarks and brief perspectives in section 8.

2. Overview and Related Works

The problem we are concerned with is multi-person (passers-by) detection and tracking in an environment co-occupied by humans (possibly crowded) and a mobile robot. This perception capability is imperative for an active robot that needs to interact with individuals in the environment. Broadly speaking, the literature in automated multi-person detection and tracking encompasses works that use sensors fixed in the environment and those that use mobile sensors (either mounted on a mobile robot or a moving vehicle). This work spans both realms by combining information from fixed sensors with information from mobile sensors. To put the proposed framework into context, it is necessary to give an overview and mention related works in: (i) fixed sensor(s) based person detection and tracking, (ii) mobile sensor(s)

based person detection and tracking, (iii) sensor fusion modes, and (iv) co-operative systems that try to combine fixed and mobile sensors.

Apparently, research works that use sensors fixed in the environment are vast in number [11, 12]; they include works that use a single classical camera, network of overlapping [12] and/or non-overlapping cameras [13, 14], and a network of heterogeneous sensors (*e.g.,* Laser Range Finders (LRFs) and vision [15]). Since the sensors are stationary, simple and fast algorithms like background subtraction and optical flow could be used to detect moving persons within the FOV. Depending on actual sensor configuration, they can encompass wide areas–therefore, provide global perception. They can view and track subjects over a broad area for an extended period of time. But, their main pitfalls include evident dead-spots that could arise from configuration (placement and number of sensors used), possible occlusions, and their passiveness.

On the other hand, mobile robot based systems, as a consequence of their mobility, are generally more suited for surveilling and/or monitoring large areas as they provide a means to reduce the environment structuring and the number of devices needed to cover a given area [16]. But, multi-person detection and tracking from mobile robots is more challenging due to on-board sensors' motion (during robot mobility), limited FOV of on-board sensors, and limited on-board computational resources. On the other hand, sensors mounted on robots provide localized perception and can pick up details. As a result, robotic based surveillance applications are mostly limited to activities that require close monitoring. They are also suitable for patrolling wide areas owing to their ability to re-position themselves. In addition, they also provide a means for action which can be of paramount advantage for following a target [5], intruder intervention [17], provision of assistance [2], and possibly physical restraint of an assailant [18].

When working with mobile robots, most researchers make use of 2D Laser Range Finders (LRFs) and vision sensors mounted extensively for human detection and tracking. 2D LRFs provide a 2D depth scan of an environment. They have high accuracy, high scanning rates, and are insensitive to lighting conditions. Since they are mostly mounted at a height corresponding to a human leg, person detection proceeds by detecting leg scan patterns in each frame [19, 20]. Some researchers have also mounted LRFs in two layers, scanning at the height of a leg and chest to improve the detection rate, *e.g.,* [21]. Unfortunately, due to their planar scan nature, they are very susceptible to occlusions and are easily fooled by geometrically leg like structures in

the environment. They are also not suitable for multi-person tracking with unique identities as they furnish no relevant information for discriminating amongst persons leading to frequent failures in crowded environments. It can be emphasized here that in these scenarios, they ought to be combined with other sensors with more rich perception capabilities. On the contrary, visual sensors provide rich information that capture persons' appearance well. To detect persons, either background subtraction techniques [22] or motion segmentation [23] can be used from a stationary mobile robot. In case of an active moving robot, recent single frame based approaches like Histogram of Orientation Gradients (HOGs) based person detection [24, 25], face detection [24], and though with questionable performance, skin color segmentation [26] can be used. For platforms equipped with stereo-vision camera, 3D human like blob segmentation is also a viable option [27]. In effect, vision based multi-person tracking implementations have shown far better results than those based on LRFs owing to rich appearance information and lessened confusion with environment structures. But, they still suffer from narrow FOVs (unless special omni-directional cameras are used), occlusions, and high processing time requirements.

Evidently, most robotic systems are equipped with various sensors and it is only natural to consider fusing the different sensor data to improve individual sensor percepts. The extent of the improvement depends on how well the different sensors complement each other. In the robotic community, fusion of LRF and vision for people detection and tracking has shown to outperform individual counterpart modalities [28, 29]. The fusion, for example, can be done in a sequential manner at the detection level, using the laser hypothesis to constrain the search in the visual data as in [25], or in the tracking step [30]. Variants of Kalman Filters [31] and Particle Filters [23] have been principally used for fusing laser and vision at the tracking step for multi-person tracking. The key interest in laser and vision fusion is combined provision of precise 3D position and rich appearance information which leads to a detection/tracking system with high precision and accuracy. The availability of wide FOV vision system further improves this performance as demonstrated through fusion of a laser with omni-directional cameras [28, 23, 32].

Furthermore, some researchers have considered fusing vision and audio data [33, 34, 35]. Audio data can be used to localize the sound source (possibly a person) and identify the speaker. These are additional features that would enrich the vision data leading to better tracking and identification in crowds. Some works have also considered special sensors like thermal

cameras [36] mounted on a mobile robot. Since humans have distinct thermal profile compared to indoor environments, they stand out bright in thermal images which leads to easy detection. But, multi-person tracking in a crowded environment using a thermal camera solely is challenging as human thermal signature is the same for every individual, leading to difficulty in tracked target discrimination amongst each other. [37] augmented a thermal camera with classical gray scale camera to realize a system that can detect individuals easily and then use the gray scale image for identification (disambiguation). Another special sensor recently burgeoning is the Kinect [38]. The Kinect provides an RGB color image and 3D information. In some works, it has been mounted on a mobile robot and used for multi-person perception by fusing the heterogeneous data it provides [6, 39]. Though highly promising, its narrow FOV still remains a problem.

Sensor fusion is certainly not limited to two sensors; depending on availability of sensors and computational time constraint, more sensor data could be fused. For example, Martin *et al.* [26] fused LRF, omni-directional camera, and a ring of sonar beams, in a probabilistic aggregation scheme to detect and track individuals in the vicinity of the robot. Zivkovic *et al.* [32] combined sensor data from an omni-directional camera, a classical camera mounted on Pan-Tilt-Unit (PTU), and LRF to detect multiple persons using a parts based model. Both cases attest that the plethora of sensors used improve performance well. The improvement comes about mainly because of the complementary nature of the utilized sensors. The rich vision information from cameras can be complemented by employing cameras with different FOVs [32],*e.g.,* wide FOV from wall mounted cameras and narrow localized FOV from a camera on a robot.

In recent years, researchers have considered surveillance systems that incorporate mobile robots and environment fixed sensors cooperatively. These cooperative surveillance systems combine the merits of fixed and mobile perception modes. They acquire global and wide area perception from the fixed sensors, localized perception and a means for action from the mobile robot. This kind of cooperative systems have the potential to lead to more generic surveillance systems as they can handle various scenarios. Li *et al.* [40] presented a time-related abnormal events detecting and monitoring system using wireless sensor network and a mobile robot. In their work, intruders are detected using the sensor networks. Upon detection, the mobile robot travels to the position to further investigate the situation locally with its camera. Similarly, in [9] three networked wall mounted fixed view cameras and a mo-

bile robot are used to track and follow a target. The target is first detected using the fixed cameras. Once detected, the information is passed onto the robot which navigates to that position and continues to follow the target person. Chakravarty *et al.* [10] presented an intruder interception system using external cameras and a mobile robot cooperatively. The external cameras are used to detect an intruder and aid the mobile robot in navigation. The mobile robot, once it has received the location of the intruder, proceeds and intercepts it acting as a means of action to the system. All the above cooperative perception systems portray similar approaches in which perception of interesting targets is initially carried out based on the fixed sensors. The mobile robot's target perception capability is delayed until target presence is communicated to the robot. The perceptual decision making is somewhat decentralized with no data fusion. There is no centralized scheme to collect evidence from the fixed and on-boarded (mobile) sensors to track the targets, rather, either the deported vision, in [10], or the mobile robot, in [9], does the tracking after the initial target detection. But, an important observation that needs to be made from the related works is data fusion actually leads to robust perception modes. Marching on this line, we propose a perceptual system that makes use of the localized perception capabilities of the sensors on the mobile robot in cooperation with external fixed cameras both to detect and distinctly track persons in the vicinity of the robot in a centralized manner. In short, a centralized data fusion between fixed sensor percepts and a mobile sensor percept is proposed. The proposed system has the ability to complement local perception with global perception and vice-versa. To the best of our knowledge this cooperative framework is unique in the literature.

Contributions: The work presented in this paper makes two core contributions, namely: (1) It proposes and validates a centralized cooperative framework and data fusion scheme between wall mounted fixed view cameras and sensors embedded on a mobile robot to track multiple passers-by in a surveilled area; this is unique in the literature as it differs from discussed existing cooperative frameworks([9, 10]). (2) It realizes a person following mobile robot system with passers-by avoidance in crowds by deploying the developed perceptual functionalities on the actual platform–with seamless, coherent integration and coupling with the robot's actuators. To the best of our knowledge, a person following mobile robot with passers-by avoidance in crowded environment does not yet exist in the literature. Even though the works of Hoeller *et al.* [41] has the capability to avoid passers-by, it has not been validated in crowded scenes and it relies only on LRF for passers-by

tracking which in our experience leads to frequent tracker failures in crowded environments. Various formulations, implementation details, and experimental validations that clearly and consistently highlight our contributions are detailed in the rest of this paper.

3. Framework and Architecture

Our goal is to realize a person following mobile robot that will follow a single target person, hereafter referred as *the user*, wearing an RFID tag while at the same time perceiving the passers-by and taking their dynamics into consideration to adjust its motion accordingly. To robustly perceive the user and all passers-by and increase the anticipation capability of our robot, we have devised a cooperative perception system made up of two fixed view wall mounted cameras and various sensors on-board a mobile robot.

3.1. Development Platform and Environment

Our cooperative framework is made up of a mobile robot and two fixed view wall-mounted RGB flea2 cameras. The cameras have a maximum resolution of 640x480 pixels and are connected to a dual-core Intel Centrino computer *via* a fire-wire cable (figure 1). The robot, called Rackham, is an iRobot B21r mobile platform. Rackham has various sensors, of which its SICK Laser Range Finder (LRF), positioned $38cm$ above the ground and with a 180^o FOV, Micropix digital camera mounted on a Directed Perception pan tilt unit (PTU), and an omni-directional RF system custom-built in the lab for detecting RF tagged person all around the robot [5], are utilized in this work. Rackham has two PCs (one mono-CPU and one bi-CPU PIII running at 850 MHz) and a Wireless Ethernet. Figure 1 shows the hardware aspect of our framework. Communication between the mobile robot and the computer hosting the cameras is accomplished through a wi-fi connection. The visual camera on Rackham has a very narrow FOV ($< 50^o$). This means the robot does not have any visual information for the rest of 310^0 of the surrounding. This limits the perception capability of the robot impairing its anticipation capability. But by fusing this information with wall mounted cameras that collectively span a wide area, this problem could be mediated if not alleviated.

Rackham's software architecture is based on the **G**eno**M** architecture for autonomy [42]. All its functionalities have been embedded in modules created by **G**eno**M** using C/C++ interface.

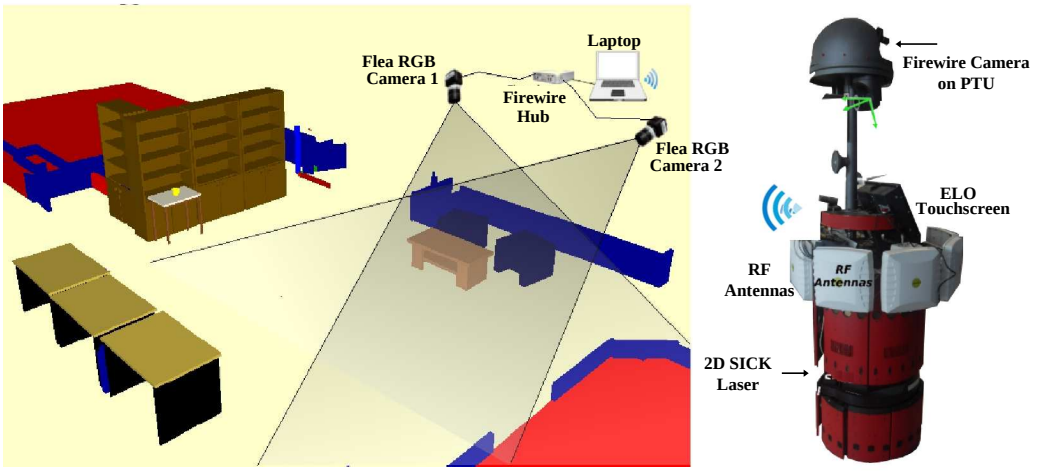


Figure 1: Perceptual platform; wall-mounted cameras (with rough positioning and fields of view) and Rackham, the mobile robot.

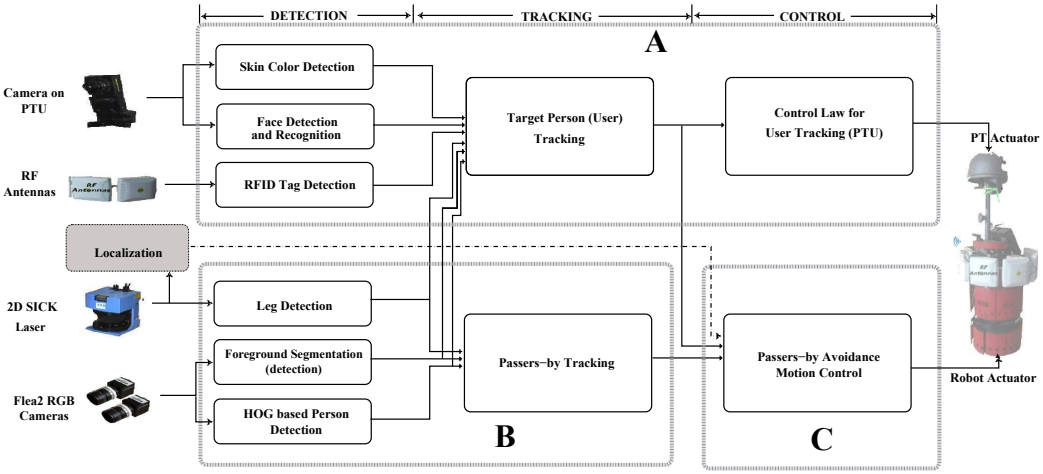


Figure 2: Overall system block diagram.

3.2. System Block Diagram and Description

A block diagram of our overall envisaged system to realize person following with passers-by avoidance with a mobile robot is shown in figure 2. It has three main components labeled A, B, and C. Block A is responsible for the user tracking and following activity and has been addressed in [5]. Briefly speaking, it relies on an RF reader and a visual camera mounted on the robot. The RF reader is capable of detecting and localizing an RFID tag

placed around the robot (360^o FOV) within the range of $0.5m$ to $4.5m$. The detection yields a unique id corresponding to the tag and a course position estimate (range and azimuth) of the tag. Only the user wears an RFID tag which plays the major role to disambiguate him/her from the passers-by. An instance of an RFID detection is shown in figure 3(c). The on-board visual sensor is used for face detection based on Viola and Jones [43] face detector and skin segmentation of close by persons. It is true a generic head detector could be used rather than a face detector here. But, we have retained the use of the face detector because of the following reasons: (i) We already have full body detections of the user from the fixed-cameras and a more complementary information would be the face; (ii) this work is in a Robotic context and further steps will involve user-robot interaction establishment which necessitates face detection to know when a person is engaging the robot (we are basically laying the ground work here); and (iii) previous experiments have been done with face recognition for identification purposes so we wanted to have this perspective open for future additions that would involve user identification based on face rather than RFID. Figure 3(b) shows an exemplar view from the on-board camera. These detections are passed onto the user (target) tracking module which fuses them to identify the user to be followed by the robotu and track it accordingly. This fusion and tracking is based on sequential Monte Carlo simulation methods. The only modification here is the addition of detections from the LRF and deported vision (presented in section 4) in the same tracking framework. The 3D target position estimate of the tracker is used as a goal to drive the robot to the person and directly control the PTU of the mobile camera to keep the target at the center of focus of the on-board vision via visual servoing techniques.

Component B is responsible for perceiving passers-by in the robot surrounding and is addressed in this paper. Component C, also addressed here, is responsible for navigation of the robot to the dynamic goal, *i.e.,* the user pursuit with passers-by avoidance based on the spatio-temporal trajectory information of humans' from the passers-by tracking modality (component B). All the sensors are connected to detection modules. The heterogeneous detection modules take the raw input from the respective sensors and automatically detect humans. The tracking modules then take the various detections as input, fuse them to make best informed user and passers-by tracks. Finally, the tracking outputs are fed to the two control laws, one controls the PTU of the visual camera on the robot to keep the target in the FOV. The second control system uses the passers-by spatio-temporal information

(a) (b) (c)

Figure 3: On-board vision and RFID detection. The view from the on-board camera, corresponding to the H/R situation in (a), is shown in (b) with a detected face. (c) shows an instance of the RFID detection. The red circle in the middle corresponds to the robot while the white circle shows the range of the RF reader. A detected tag is shown as a saliency map in this region depending on its azimuth and range. The red arc shows the pan position of the on-board camera.

to avoid the passers-by while following the user.

4. Multi-person Detection

The perceptual functionalities of the entire system are based on various detections. The detection modules are responsible for automatically detecting persons in the area. Different person detection modalities are utilized depending on the data provided by each sensor.

4.1. Leg Detection with LRF

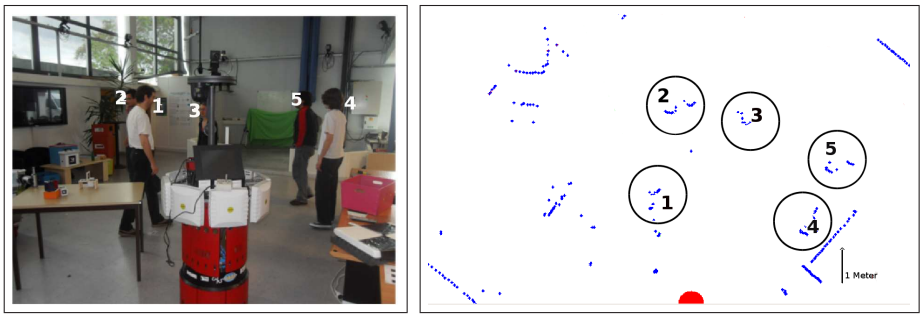
Laser Range Finders (LRFs) have become attractive tools in the robotics area for environment detection due to their accuracy and reliability. As the LRFs rotate and acquire range data, they will have distinct scan signatures corresponding to the shape of an obstacle in the scan region. In our case, the LRF provides horizontal depth scans with a 180^o FOV and 0.5^o resolution at a height of $38cm$ above the ground. Person detection, hence, follows by segmenting leg patterns within the scan. In our implementation a set of geometric properties characteristic to human legs outlined in [44] are used. Figure 4 shows an instance of a scan with leg signatures circled and the actual human-robot situation. The detection proceeds in three steps:

1. *Blob segmentation.* All sequential candidate scan points that are close to each other are grouped to make blobs of points. The grouping is done based on the distance between consecutive points.

2. *Blob filtering.* The blobs formed are filtered using geometric properties outlined in [44]. The filtering criteria used are: *Number of scan points*, *Mid point distance*, *Mean Internal Angle and Internal Angle Variance*, and *Sharp structure removal*. For details on these criteria, the reader is referred to [44].

3. *Leg formation.* All the blobs that are not filtered out by the above stated requirements are considered to be legs. Each formed leg is then paired with a detected leg in its vicinity (if there is one). The center of the paired legs makes the position of the detected human.

Each person detection has an associated appearance representation obtained by projecting a rectangular region, corresponding to an average person, onto the wall mounted camera images thanks to the fully calibrated system. The appearance is captured in the form Hue-Saturation-Value (HSV) [45] histogram. Individual histograms are obtained from the two cameras, of course if the detections are within the field of view, and are treated separately. These detections are passed on to both the passers-by tracking and user (target) tracking modules.



(a) (b)

Figure 4: LRF scan illustrations showing the human-robot situation in (a) and the associated laser scan in (b). Scans corresponding to legs are shown circled. Rackham is shown as the red circle in (b).

4.2. Foreground Segmentation (Detection)

The two wall mounted cameras with partially overlapping FOV provide a video stream of the area. One person detection mode employed is foreground segmentation using background subtraction as these cameras are static. To

accomplish this, a simple Σ-Δ background subtraction technique [46] is used. After a series of morphological operations, only foreground blobs with an aspect ratio comparable to an average human being are kept and treated as detected persons. The mobile robot is masked out of the foreground images using its position from its localization module. The person detections are matched with an adaptive HSV appearance model of *the user*. Detections very similar to the model, only one per camera, are assumed to be of *the user* and are passed along to the target tracking module whereas the rest are passed along to the passers-by tracking module. The target appearance model is initially initialized with the appearance histogram captured during the beginning of the person following activity and is distinct to each deported camera. The rest of the detections are projected to yield ground positions, $(\mathsf{x}, \mathsf{y})_G$, with associated color appearance information (in the form of HSV histograms) of individuals in the area and are then passed along to the passers-by tracking module. Figure 5b shows sample foreground segmented image with bounding box to show detected humans from both cameras.

(a) (b) (c)

Figure 5: Sample images from the two wall mounted cameras. (a) shows the actual feed, (b) shows the segmented foreground/background image based on Σ-Δ background subtraction technique with bounding boxes, and (c) shows HOG based detections.

4.3. HOG Person Detection

Similar to the foreground segmentation step, HOG person detection [47] is used to automatically detect persons in the surveilled area using the feed

from the wall mounted cameras. This method makes no assumption of any sort about the scene or the state of the camera (mobile or static). It detects persons in each frame using histogram of orientation gradient features. Again, detections very similar to the appearance of the target person, only one per camera, are passed to the user tracking module. The rest are passed to the passers-by module once projected into ground position, $(\mathsf{x}, \mathsf{y})_G$, with an associated HSV histogram. Sample HOG based person detections are shown in figure 5(c); corresponding sample HSV histograms computed as in [45] are shown in figure 6. The histograms have an 8x8 HS bin and 8 V bin. They are shown unrolled in a single dimension for ease of visualization.

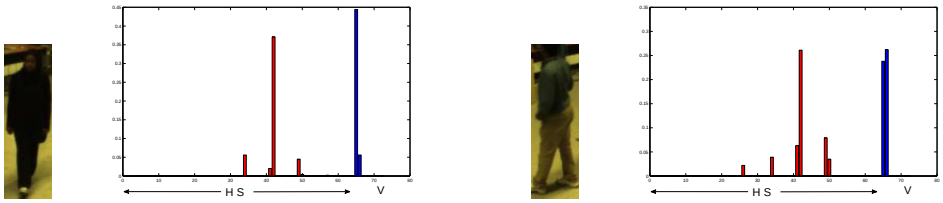


Figure 6: HSV histograms computed for two targets. The histograms have 8x8 HS bin and 8 V bin. They are shown here unrolled in a single dimension.

In summary, five sub-set of detections are produced for the passers-by tracking module, namely: one from the LRF (l), two from the wall mounted cameras via foreground segmentation ($fseg_{c_1}, fseg_{c_2}$), and another two via HOG detection from the same cameras (hog_{c_1}, hog_{c_2}). Hence, the complete set of detections passed along at time t is denoted as $\{z^d_{t,j} : d \in \{l, fseg_{c_1}, fseg_{c_2}, hog_{c_1}, hog_{c_2}\}, j \in \{1, .., N_d\}\}$ where N_d represents the number of detections in the d^{th} detector and each z denotes a detected person position on the ground floor $(\mathsf{x}, \mathsf{y})_G$.

5. Passers-by Tracking

Tracking of passers-by is the problem of Multi-Person Tracking (MPT), which is concerned with the problem of tracking a variable number of persons–possibly interacting. Our aim here is to correctly track and obtain trajectories of multiple persons within the field of view of the utilized sensors. The literature in multi-target tracking contains different approaches. Multiple Hypothesis Tracker (MHT)[48], Joint Probabilistic Data Association Filter

(JPDAF)[49], centralized [50] and decentralized Particle Filters (PFs) [51], and Markov Chain Monte Carlo Particle Filtering (MCMC-PF) [52]. MHT is computationally expensive as the number of hypothesis grows exponentially over time, while JPDAF is applicable to tracking a fixed number of targets. The decentralized particle filtering scheme, based on multiple independent PFs per target, suffers from the "hijacking" problem since whenever targets pass close to one another, the target with the best likelihood score takes the filters of nearby targets. The centralized PF scheme–a particle filter with a joint state space of all targets–is not viable for more than three or four targets due to the associated computational requirement. A more appealing alternative in terms of performance and computational requirement is the MCMC-PF. MCMC-PF replaces the traditional importance sampling step in joint PFs by an MCMC sampling step overcoming the exponential complexity and leading to a more tractable solution. For varying number of targets, Reversible Jump Markov Chain Monte Carlo - Particle Filters (RJMCMC-PFs), an extension of MCMC to variable dimensional state space, has been pioneered to perform successful tracking [52]. When it comes to tracking multiple interacting targets of varying number [52] has clearly shown that RJMCMC-PFs are more appealing taking performance and computational requirements into consideration. This is also attested in various recent research works,*e.g.,* [6, 53, 54]. Inspired by this, we have used RJMCMC-PF, adapted to our cooperative perceptual strategy, for passers-by tracking driven by the various heterogeneous detectors.

5.1. RJMCMC-PF Formalism

RJMCMC-PF replaces the importance sampling step of Particle Filters with an RJMCMC sampling step. For a given sequence of measurements upto a time t, denoted as $Z_{1:t}$, the posterior distribution over the targets state, $P(X_t|Z_{1:t})$, is approximated in terms of N unweighted samples, $P(X_t|Z_{1:t}) \approx \{X_t^n\}_{n=1}^N$ where X_t^n denotes the n^{th} particle, in a Bayesian Framework. The state of a particle in RJMCMC-PF encodes the configuration of the entire tracked targets: $X_t^n = \{I_t^n, x_{(t,i)}^n\}, i \in \{1, ..., I_t^n\}$, where I_t^n is the number of tracked objects of hypothesis n at time t, and $x_{(t,i)}^n$ is a vector encoding the state of object i. The posterior estimation is achieved by defining a Markov Chain over the variable dimension state space configuration X_t^n such that the stationary distribution of the chain approximates the desired posterior well. Roughly, at each time step, the filter starts the Markov Chain from a sampled initial configuration and iterates $N + N_B$ times, where N is

the number of particles and N_B represents the number of burn-in iterations needed to converge to stationary samples. At time t, in each iteration,n, of the RJMCMC-PF, the filter proposes a new state hypothesis, X^*, from the previous iteration state hypothesis (X^* from X_t^{n-1}) depending on the chosen proposal move that either varies or leaves the dimension of the state unaltered. The final N particles represent the sought approximation to the required posterior.

5.2. Implementation

Our RJMCMC-PF tracker is driven by the heterogeneous detectors that provide ground position of individual persons and their corresponding appearance information (section 4). The actual detectors are: the LRF based person detector, the foreground segmentation (detection) from each wall mounted camera, and the HOG based person detector on each wall mounted camera. The passers-by tracking is performed on the ground plane. RJMCMC-PF accounts for the variability of the tracked targets by defining a variable dimension state space. The state space dimension is considered as a union of several subspaces. Whenever tracking of a new passer-by starts, the state "jumps" to a large dimensional subspace and there will be a "jump" to a low dimensional subspace whenever a tracked person is removed from a hypothesis. In each iteration, the state space exploration is driven by the proposal move q_m that proposes a specific move and computation of the acceptance ratio β according to the chosen move. Equation 1 shows computation of the acceptance ratio of a proposal X^* at the n^{th} iteration. It makes use of the jump move distribution, q_m; proposal move distribution, $Q_m()$, associated with each move; the observation likelihood, $\pi(X_t^n)$; and the interaction model, $\Psi(X_t^n)$. Our choice of these components that are crucial to any RJMCMC-PF implementation are briefly discussed below. The complete passers-by tracking implementation is summarized in algorithm 1.

$$\beta = \min\left(1, \frac{\pi\left(X^*\right) Q_{m*}\left(X_t^{n-1}|X^*\right) q_{m*}\Psi\left(X^*\right)}{\pi\left(X_t^{n-1}\right) Q_m\left(X^*|X_t^{n-1}\right) q_m\Psi\left(X_t^{n-1}\right)}\right) \tag{1}$$

where m denotes the proposed move and m^* denotes the reverse move.

5.2.1. State space

The state vector of a person i in hypothesis n at time t is a vector encapsulating the id and $(\mathsf{x}, \mathsf{y})_G$ position of an individual on the ground plane with

Algorithm 1: RJMCMC-PF Passers-by Tracking

input : $\{X_{t-1}^n\}_{n=1}^N$; $\hat{X}_{t-1}$; $\{z_t\}$
output: $\{X_t^n\}_{n=N_B}^{N+N_B}$ and $\hat{X}_t$ (MAP estimate)

Init: *pick a random particle from the set* $\{X_{t-1}\}$ *with similar configuration to* $\hat{X}_{t-1}$ *and perturb each target with a zero mean Gaussian to obtain* X_t^0;

for $n \leftarrow 1$ **to** $N + N_B$ **do**

Choose a move $m \in \{$`Add`, `Update`, `Remove`, `Swap`$\} \sim q_m$;

switch m **do**

case `Add`:

$X^* = \{X_t^{n-1}, x_p\}$; x_p *is randomly taken from* $\{z_{t,j}^d\}$;

$\beta = \min\left(1, \frac{\pi(X^*)Q_{remove}\left(X_t^{n-1}|X^*\right)q_{remove}\Psi(X^*)}{\pi\left(X_t^{n-1}\right)Q_{add}\left(X^*|X_t^{n-1}\right)q_{add}\Psi\left(X_t^{n-1}\right)}\right)$; break;

case `Remove`:

$X^* = \{X_t^{n-1} \setminus x_p\}$ *where* $p \in \{1, ..., I^{n-1}\}$;

$\beta = \min\left(1, \frac{\pi(X^*)Q_{add}\left(X_t^{n-1}|X^*\right)q_{add}\Psi(X^*)}{\pi\left(X_t^{n-1}\right)Q_{remove}\left(X^*|X_t^{n-1}\right)q_{remove}\Psi\left(X_t^{n-1}\right)}\right)$;

break;

case `Update`:

Randomly select a target x_p *from* X_t^{n-1};

Select x_p^*, *a random subspace corresponding to* x_p *in the particle set at* $t-1$;

Replace x_p *with a sample from* $\mathcal{N}\left(; x_p^*, \Sigma\right)$ *proposing* X^*;

$\beta = \min\left(1, \frac{\pi(X^*)\Psi(X^*)}{\pi\left(X_t^{n-1}\right)\Psi\left(X_t^{n-1}\right)}\right)$; break;

case `Swap`:

Swap the ids of two near tracked persons to propose X^*;

$\beta = \min\left(1, \frac{\pi(X^*)\Psi(X^*)}{\pi\left(X_t^{n-1}\right)\Psi\left(X_t^{n-1}\right)}\right)$; break;

if $\beta \geq 1$ **then** $X_t^n \leftarrow X^*$;

else

Accept $X_t^n \leftarrow X^*$ *with probability* β *or reject and set* $X_t^n \leftarrow X_t^{n-1}$;

Discard the first N_B *samples of the chain*;

MAP estimate, $\hat{X}_t := E_{p(X_t|Z_{1:t})}[X_t] = \arg\max_{X_t^n}[count(x_k^n)]$;

respect to a defined coordinate base, $x_{t,i}^n = \{Id_i, \mathtt{x}_{t,i}^n, \mathtt{y}_{t,i}^n\}$. Consequently, the n^{th} particle at time t is represented as $X_t^n = \{I_t^n, x_{(t,i)}^n\}, i \in \{1, ..., I_t^n\}$, where I_t^n is the number of tracked persons by this particle at time t.

5.2.2. Proposal moves

Four sets of proposal moves are used: $m = \{\mathtt{Add}, \mathtt{Update}, \mathtt{Remove}, \mathtt{Swap}\}$. The choice of the proposal privileged in each iteration is determined by q_m, the jump move distribution. These values are determined empirically and are set to $\{0.15, 0.8, 0.02, 0.03\}$ respectively. They are tuned to better reflect the occurrences of these events in the scene. It is evident that, once a target appears in the scene, he/she does not disappear immediately. So there will more `Update` moves rather than `Add`, `Remove`, and `Swap` moves. These values could actual be set arbitrary, but then a lot of MCMC iterations would be required to obtain a steady state approximation of the posterior. To formulate the proposal move distributions, $Q_m()$, a Gaussian Mixture model is used. A Gaussian distribution better represents the confidence obtained from a detector and tracker that provides a point estimate for the target position. This distribution clearly exemplifies the highest confidence at the point estimate (mean) and how the confidence wears off as we move away from the centroid radially.

To simplify both the transition of the new proposed state hypothesis X^* (at the n^{th} iteration from X_t^{n-1} at time t) and evaluation of the acceptance ratio only changes to a randomly chosen subset of the state is considered. In multi-target tracking, this translates into changing a single target per iteration.

Add: The add move, randomly selects a detected person, x_p, from the pool of provided detections and appends its state vector on X_t^{n-1} resulting in a proposal state X^*. The proposal density driving the `Add` proposal, $Q_{Add}\left(X^*|X_t^{n-1}\right)$, is then computed according to equation 2. This equation represents a mixture of Gaussian map made from the detected passers-by and tracked passers-by at time $t-1$. Each detection is represented as a Gaussian on the ground plane. It is then masked by a similar mixture derived from the tracked persons (Maximum A Posteriori (MAP) estimate $\hat{X}$) at time $t-1$ in such a way that the distribution will have higher values on locations conforming to detected passers-by that are not yet being tracked. The covariance matrix used for all Gaussian mixtures from detector and tracking are identical to simplify normalization.

$$Q_{add}\left(X^{*}|X_{t}^{n-1}\right)=\sum_{d}\frac{k_{d}}{N_{d}}\sum_{j=1}^{N_{d}}\mathcal{N}\left(x_{p};z_{t,j}^{d},\Sigma\right).\left(1-\frac{1}{N_{T}}\sum_{j=1}^{N_{T}}\mathcal{N}\left(x_{p};\hat{X}_{t-1,j},\Sigma\right)\right) \tag{2}$$

Where d represents the set of detectors, namely: from laser (l), fixed camera 1 (c_1), and fixed camera 2 (c_2); $d \in \{l, c_1, c_2\}$ (each camera has two detections HOG, hog, and Foreground Segmentation, $fseg$), N_d is the total number of detections in each detector, k_d is a weighting term for each detector such that $\sum_d k_d = 1$, $\hat{X}_{t-1}$ is the MAP estimate of the filter at time $t-1$, and N_T is the number of targets in this MAP. Figure 7 clearly illustrates what the add move proposal density looks like on a specific situation. When a new passer-by is added, its appearance is cross-checked with the appearance of passers-by that have been tracked. If there is a high similarity, determined based on Bhattacharyya distance, the new person is given the id of the matched person and the situation is treated as a simple re-identification step.

Remove: The remove move, randomly selects a tracked person x_p from the particle being considered, X_t^{n-1}, and removes it, proposing a news state X^*. Contrary to the add move, the proposal density used when computing the acceptance ratio, $Q_{Remove}(X^*|X_t^{n-1})$ (equation 3), is given by the distribution map from the tracked persons masked by a map derived from the detected passers-by. This distribution favors removal of targets that have gone out of the tracking area but are still being tracked.

$$Q_{remove}\left(X^{*}|X_{t}^{n-1}\right)=\left(1-\sum_{d}\frac{k_{d}}{N_{d}}\sum_{j=1}^{N_{d}}\mathcal{N}\left(x_{p};z_{t,j}^{d},\Sigma\right)\right).\left(\frac{1}{N_{T}}\sum_{j=1}^{N_{T}}\mathcal{N}\left(x_{p};\hat{X}_{t-1,j},\Sigma\right)\right) \tag{3}$$

Even though the tracker of a person who left the scene ceases to exist, a dynamic appearance model of the person is kept for a later re-identification.

Update: In the update proposal move, the state vector of a randomly chosen passer-by is perturbed by a zero mean normal distribution. The update proposal density, $Q_{update}(X^*|X_t^{n-1})$, is a normal distribution with the position of the newly updated target as mean. Hence, the acceptance ratio

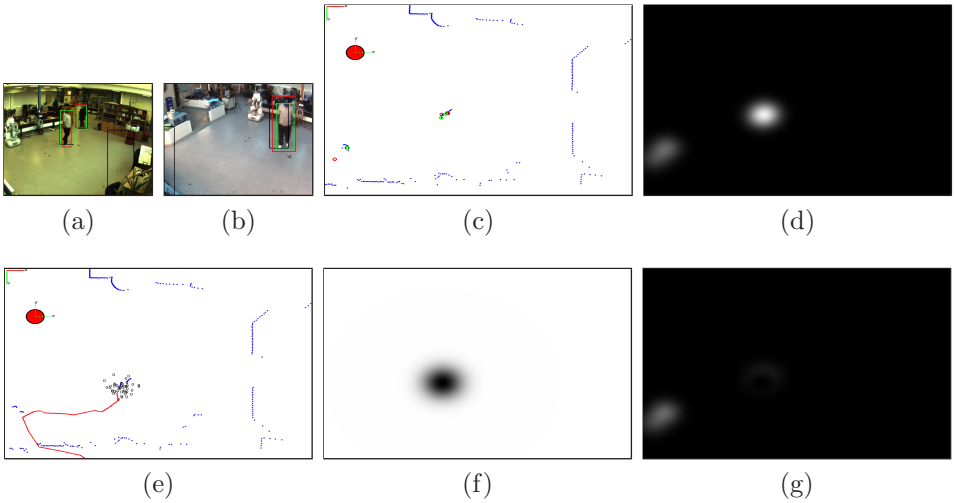


Figure 7: Illustration of the add proposal distribution. (a)-(b) shows the wall mounted cameras feed with various detections (laser in black, foreground segmentation in red, and HOG in green). (c) projection of each detection onto the ground plane. (d) shows the mixture of Gaussian distribution determined from the various detections. (e) shows the tracked target at time $t-1$ and (f) shows its corresponding Gaussian mask. Finally, (g) shows the add proposal distribution obtained by masking (d) with (f), and it indeed shows salient values on the position of the untracked passer-by.

is influenced only by the likelihood evaluation and interaction amongst the targets.

Swap: The swap move handles the possibility of id switches amongst near or interacting targets. When this move is selected, the ids of the two nearest tracked persons are swapped and a new hypothesis X^* is proposed. The acceptance ratio is computed similar to the `Update` move.

5.2.3. Interaction model ($\Psi(.)$)

Since the passers-by are likely to interact, an Interaction Model is included to maintain tracked person identity and penalize fitting of two trackers to the same object during interaction. Similar to [52, 55], a Markov Random Field (MRF) is adopted to address this. A pairwise MRF where the cliques are restricted to the pairs of nodes (targets define the nodes of the graph), that are directly connected to the graph, is implemented as part of our tracker. For a given state X_t^n, the MRF model is given by equation 4. As can be seen from this equation, as long as the σ term is not set to zero, $\phi(,)$ will always

be greater than 0 and less than 1. The sigma determines how well the effect should be pronounced when the targets are close by.

$$\Psi(X_t^n) = \Pi_{i \neq j} \phi(x_{t,i}^n, x_{t,j}^n)$$
$$\phi(x_{t,i}^n, x_{t,j}^n) = 1 - exp\left(-\left(\frac{d(x_{t,i}^n, x_{t,j}^n)}{\sigma}\right)^2\right) \quad (4)$$

where $d(x_{t,i}^n, x_{t,j}^n)$ is Euclidean distance; $i, j \in \{1, ..., I_t^n\}$; and I_t^n is the number of targets in X_t^n.

5.2.4. Observation Likelihood ($\pi(.)$)

The observation likelihood, $\pi(.)$ in equation 1, is derived from all detector outputs except the LRF for which blobs formed from the raw laser range data are considered. If the specific proposal move is an `Update` or `Swap` move, a Bhattacharyya likelihood measure is also incorporated. The raw laser data is filtered to make blob and keep those within a range of radius, denoted as l_b. This filters out laser data pertaining to walls, thin table or chair legs, and other wide structures. Then every filtered blob is represented as a Gaussian on the ground plane centered on the centroid of the blob. HOG based person detection, and detection from foreground segmentation are also represented as a Gaussian mixtures on the ground plane averaged over the number of detections with each Gaussian centered on the detection points. Representing the measurement information at time t as z_t, the observation likelihood of the n^{th} particle X_t^n at time t is computed as shown in equation 5.

$$\pi\left(X_t^n\right) = \pi_B\left(X_t^n\right).\pi_D\left(X_t^n\right) \quad (5)$$
$$\pi_B\left(X_t^n\right) = \begin{cases} \prod_{i=1}^{M} \prod_{c=1}^{2} e^{-\lambda B_{i,c}^2}, & \text{if } move = \texttt{Update} \text{ or } \texttt{Swap} \\ 1 & , \quad \text{otherwise} \end{cases}$$
$$\pi_D\left(X_t^n\right) = \frac{1}{M} \sum_{i=1}^{M} \left(\sum_d k_d.\pi\left(x_i | z_t^d\right)\right), \sum_d k_d = 1$$
$$\pi\left(x_i | z_t^d\right) = \frac{1}{N_d} \sum_{j=1}^{N_d} \mathcal{N}\left(x_i; z_{t,j}^d, \Sigma\right)$$

In equation 5, $B_{i,c}$ represents the Bhattacharyya distance computed between the appearance histogram of a proposed target i in particle X_t^n and the target model in each camera c. M represents the number of targets in the particle, and N_d the total number of detections in each detection modality d, $d = \{l_b, c_1, c_2\}$, in this case including the measures from the laser blobs. k_d is a weight assigned to each detection modality taking their respective accuracy into consideration and x_i represents the position of target i in the ground plane.

At this point, it is interesting to point out that, even though the fusion of information from only three sensors (laser and two wall mounted cameras) is considered, the framework is equally applicable for the fusion of more heterogeneous sensors.

5.2.5. Adaptive Color Appearance Model

For each tracked passer-by, an adaptive color appearance model in the form of an HSV histogram per camera, h_{id}^c, is stored. This histogram is kept even after passers-by have left the scene. It is mainly used to re-identify a previously tracked passer-by when a new track is initiated on him/her. The new track could be initiated either due to re-entrance of the passer-by in the surveiled arena once having left, or re-initialization after tracker failure. Whenever a new passer-by is added, its color histogram is checked with existing models. If the Bhattacharyya distance is below a threshold value β_o, the new track is given the id corresponding to the matched histogram. In each time step, the appearance model of tracked passers-by is updated according to equation 6 only if the Bhattacharyya distance with the adaptive model and estimated target histogram is below a threshold value β_t.

$$h_{id}^c(t) = \alpha * h_{id}^c(t-1) + (1-\alpha) * \hat{h}_{id}^c(t) \tag{6}$$

Where $h_{id}^c(t)$ represents the adaptive histogram of passer-by id in the camera c at time step t, and $\hat{h}_{id}^c$ corresponds to the current passer-by's appearance computed at the estimated position. α is a weighing term that determines how much the current appearance affects the global model.

6. Offline Evaluations

To evaluate the performance of our RJMCMC-PF multi-person tracker, three sequences acquired using Rackham (kept static during acquisition)

and the wall mounted cameras are used. The sequences are acquired inside LAAS's robotic room which has an area of approximately 10x8.20m^2 where Rackham can actually move. Each sequence contains a laser scan and video stream from both cameras. Sequence I is a 200 frame sequence containing two targets in each frame. Similarly, sequence II contains 200 frames featuring three moving targets. Sequence III contains four targets moving in the vicinity of the robot and is 186 frames long. The quantitative performance evaluation is carried out using the CLEAR MOT metrics [56] which are the de-facto for evaluating multi-object tracking.

The following metrics are computed and reported:

- Tracking Success Rate (TSR): given by $\frac{1}{J_T}\sum_{k,j}\delta_{k,j}$ where $\delta_{k,j} = 1$ if target j is tracked at frame k, else 0. $J_T = \sum_{k,j} j_k$, and j_k represents the number of persons in the tracking area at frame k.

- Miss Rate (MR): is the ratio of misses in the sequence, computed over the total number of objects in all frames, i.e. $\frac{1}{J_T}\sum_{k,j}\delta_{k,j}$ with $\delta_{k,j} = 1$ if the target j in the area is not tracked by any tracker at frame k, else 0.

- Ghost Rate (GR): computes the number of candidate targets over no target (ghosts) averaged over the total number of targets in the dataset, i.e. $\frac{1}{J_T}\sum_{k,j}\delta_{k,j}$ with $\delta_{k,j} = 1$ if tracked target j is a ghost at frame k, else 0.

- Mismatch: mismatch error occurs when an existing tracked target is initialized as a new target or takes the id of another existing tracked target. mismatch is computed by averaging the number of mismatch errors over the total number of targets in the dataset.

- Multiple Object Tracking Precision (MOTP): measures how precisely the targets are tracked as the sum of the error between tracker position estimate and ground truth averaged over the total number of correct tracks made. It is expressed in centimeters (cms).

- Multiple Object Tracking Accuracy (MOTA): is an accuracy metrics computed by taking the total errors (Miss Rate, False Positive (FP), and Mismatch) in each frame k into consideration.

$$MOTA = 1 - \frac{\sum_k (MR_k + FP_k + Mismatch_k)}{J_T} \tag{7}$$

- Id Swap: this criterion quantifies how many times an id switch between two different tracked targets occurred. It is represented as $\sum_k \sum_{i,j} \delta_{i,j}$, where $\delta_{i,j} = 1$ when an id switch occurs between tracked target i and j in frame k, otherwise it is 0,

In the above criteria, an observation that should be made is the delineation of Mismatch and Id Swap. The Mismatch criterion counts the number of mismatches that occur for all tracks. It counts both initialization of an already tracked target with a new identifier and id swap between tracked targets as a mismatch error. For our application, id swaps are very detrimental; but, if a target is initialized as a new target, the error imparted is less severe. Hence, id swaps are separately reported in table 6.

For evaluation, a hand labeled ground truth with (x, y) ground position and unique id for each person is used for each sequence. A person is considered to be correctly tracked (True Success), if the tracking position is within a 30 cm radius of the ground truth. Each sequence is run eight times to account for the stochastic nature of the filter. Results are reported as mean value and associated standard deviation. The values set for various parameters (determined empirically) to produce the evaluation results reported in this section are listed in table 1.

Table 1: Parameter values used to produce the results reported in this section.

Symbol	Stands for	Value
k_d	detector weights, $d = \{l, c_1, c_2\}$ with $c_i = \{fseg_{ci}, hog_{ci}\}$	$k_d = \{0.16, \{0.22, 0.2\}, \{0.22, 0.2\}\}$
q_m	jump move distribution	$q_m = \{0.15, 0.8, 0.02, 0.03\}$
Σ	random walk and Gaussian mixture covariance (m^2 units)	$\begin{bmatrix} 0.09 & 0 \\ 0 & 0.09 \end{bmatrix}$
σ	interaction model standard deviation (cm)	75 cms
N	number of particles in RJMCMC-PF	150
N_B	number of burn-in iterations in RJMCMC-PF	40
HSV bins	color histogram bins	8x8 HS bin, 8 V bin
β_t	passer-by appearance model update threshold	0.24
β_o	threshold for conclusive similarity of a new passer-by with existing color model	0.1
α	passers-by dynamic color model update weight	0.9

To clearly highlight the advantage of using each sensor, the passers-by tracker is evaluated based on the following different modes:

1. Passers-by tracking using LRF input only. Results are reported in table 2. In this case, a detector weight of 1.0 is used for the laser and zero for the rest.

2. Passers-by tracking using the wall mounted cameras only. Similarly, results are reported in table 3. A detector weight of 0.5 is used for each camera with equally influential HOG and foreground segmentation detections and zero for the laser.

3. Cooperative passers-by tracking using a single camera and LRF. The results pertaining to this evaluation mode are reported in table 4. The corresponding detector weight used is a 0.4 for the laser and a 0.6 for the camera.

4. Complete system, passers-by tracking using the two wall-mounted cameras and LRF. Results are reported in 5. The detector weight parameters reported in table 1 are used.

Table 2: Laser-based only perception

Sequence	TSR		MR		GR		Mismatch		**MOTP**		**MOTA**	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
I	0.757	0.034	0.252	0.034	0.396	0.042	15.00	2.618	15.62	2.340	0.410	0.049
II	0.667	0.033	0.333	0.033	0.527	0.104	21.62	5.450	19.90	1.664	0.273	0.068
III	0.606	0.044	0.394	0.044	0.541	0.103	46.75	4.921	21.94	1.745	0.202	0.068

Table 3: Wall-mounted cameras-based only perception

Sequence	TSR		MR		GR		Mismatch		**MOTP**		**MOTA**	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
I	0.897	0.006	0.103	0.006	0.087	0.034	7.60	1.817	19.80	0.140	0.797	0.025
II	0.817	0.049	0.182	0.048	0.089	0.017	19.17	3.920	22.79	1.350	0.708	0.05
III	0.734	0.050	0.265	0.050	0.248	0.016	57.60	14.15	28.44	1.601	0.4588	0.067

Table 4: Cooperative perception using a single wall-mounted camera

Sequence	TSR		MR		GR		Mismatch		**MOTP**		**MOTA**	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
I	0.932	0.023	0.068	0.023	0.110	0.014	1.333	1.633	17.52	1.80	0.825	0.030
II	0.859	0.032	0.140	0.032	0.147	0.030	10.50	4.680	17.63	1.643	0.713	0.055
III	0.725	0.037	0.274	0.037	0.339	0.069	47.40	6.986	22.83	1.00	0.402	0.051

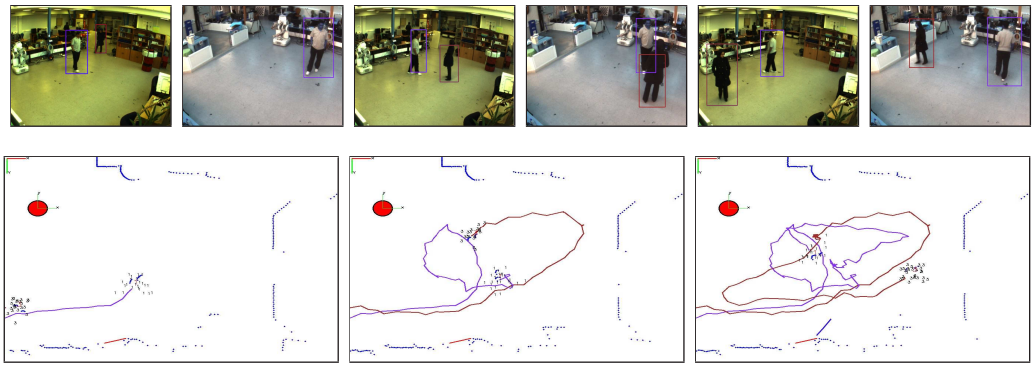
The results presented from table 2 to table 6 clearly attest the improvements in perception brought by the cooperative fusion of LRF and wall

Table 5: Cooperative perception using the two wall-mounted cameras

Sequence	TSR		MR		GR		Mismatch		**MOTP**		**MOTA**	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
I	0.935	0.029	0.065	0.022	0.099	0.020	0.667	0.816	17.01	1.886	0.841	0.033
II	0.885	0.029	0.115	0.029	0.099	0.020	11.40	3.782	17.73	0.005	0.793	0.030
III	0.755	0.018	0.245	0.018	0.211	0.027	35.60	5.941	21.30	1.358	0.538	0.040

Table 6: Id swap occurrences in each tracking mode.

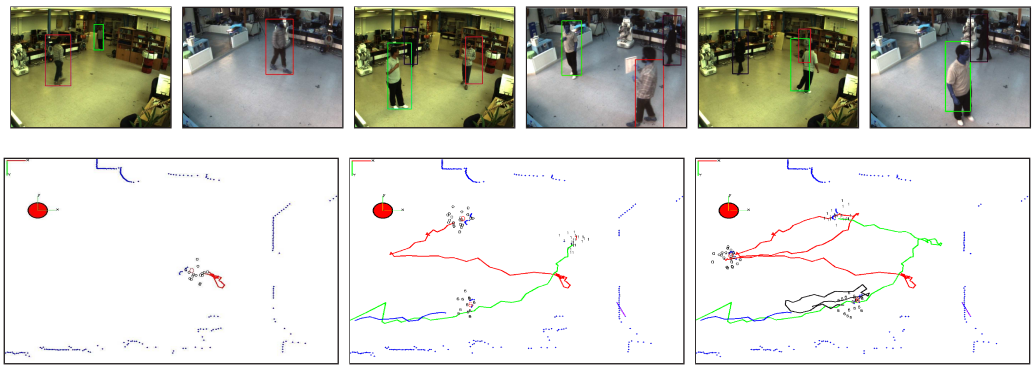
Sequence	LRF-only		Wall-mounted cameras		Cooperative Perception			
					Single Camera		Two Cameras	
	μ	σ	μ	σ	μ	σ	μ	σ
I	2.50	0.76	0.60	0.55	0.00	0.00	**0.00**	**0.00**
II	4.62	0.74	1.33	0.52	0.83	0.41	**0.40**	**0.55**
III	4.88	1.35	2.40	0.55	1.60	0.89	**1.20**	**1.09**



(a) (b) (c)

Figure 8: Multi-person tracking illustrations taken from sequence I at a) frame 31, b) frame 80, and c) frame 148. The top two images correspond to the camera streams and the bottom one shows the ground floor with trajectories of tracked persons superimposed. The particle swarm is also shown with the ID of each individual. The small blue dots are the raw laser scan points.

mounted camera percepts. The cooperative system consisting of LRF and two wall mounted cameras exhibit an MOTA of 0.841 when tracking two targets, 0.793 for three targets, and 0.538 for four targets with a 93.4%, 88.5%, and 75.5% True Success Rates respectively. The worst average precision is less than 22cms. These results are clearly indicative of how well the system does. Sample tracking sequences for two targets and three targets are shown



(a) (b) (c)

Figure 9: Multi-person tracking illustrations taken from sequence II at a) frame 27, b) frame 60, and c) frame 94. The top two images correspond to the camera streams and the bottom one shows the ground floor with trajectories of tracked persons superimposed.

in figures 8 and 9 consecutively [1]. Another main observation to make is the low accuracy of tracking based on LRF only. The mistakes made with leg like structures in the environment, sensitivity to occlusion, and lack of discriminating information amongst tracked passers-by corroborate to the poor performance. The results obtained using the wall mounted cameras show major improvements though their position tracking precision is relatively lower compared to those which include LRF measurement. By comparing table 4 and 5, it is possible to observe that the addition of the second camera in the cooperative scheme improves the tracking results further.

In table 6, id swap occurrences under each mode are reported. As specified earlier, this quantity is important to our system because identifier swaps would lead to a false motion estimation which in turn would affect the navigation of the mobile robot. Amongst the reported modes, LRF only tracking does worst. This is again expected as no appearance information to identify one person from another. Hence, LRFs should be used in conjunction with another sensor that has discriminating information where ever possible. Again, the cooperative system with two cameras results in the best results, with almost no id swaps when tracking two and three targets, and 1 to 2 id swaps with four targets through out the sequence.

Two main conclusions can generally be drawn from the results reported

[1]The reader is referred to the URL homepages.laas.fr/aherbulo/videos/cviu/ for complete runs.

in this section. First, classical video-surveillance approaches that rely on fixed visual sensors improved the perception capability of a mobile sensor unit (in our case a robot). The improvement clearly comes from the cameras that provide rich global wide field of view feed to the robot. For the second case, lets consider the evaluation that uses only deported cameras. This system is basically the same as a typical video-surveillance system made up of two networked cameras. The algorithms that we have proposed and implemented are variants of currently considered state-of-the-art algorithms. But, these results were further improved by the addition of a mobile sensor unit. Hence, it can be claimed indoor video-surveillance systems can be generally improved with a mobile sensor unit which on top of everything is also a means for action.

In short, even though the actual reported results depend on the used environment structure, it is clear that the fusion of heterogeneous sensors cooperatively increases the performance consistently. On another note, the implemented passers-by tracking has some pitfalls. Its first shortcoming comes from a formulation inherent in the RJMCMC-PF. The interaction model in this tracker depends on the state-space of the particles and not on the observations. It relies on the inference rather than the evidence. The second shortcoming relates to the employed simple persons' appearance model. Whenever a track fails (loses its target), a new track is initialized after cross-checking the appearance with past tracked targets. If this appearance is not very discriminative, it could lead to a new track initialization rather than assigning the lost track to the current target. Briefly, the simple histogram based appearance model used could easily confuse persons with similar clothing and lead to erroneous interpretations.

7. Passers-by Avoidance and Live Robotic Demo

To realize the passers-by avoidance control law, we formulated the parameterized security zone shown in figure 10 around each passer-by, along his/her travel direction inferred from the passers-by tracking system. This zone, parameterized by r, R, and ω, ensures the robot crosses the person's trajectory behind rather than in front, avoiding interference. With this representation, Rackham's basic obstacle avoidance module based on the classical Nearness Diagram (ND) Navigation [57] is used with the followed user's position, obtained from the user tracking module, defined as the goal.

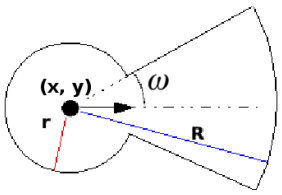


Figure 10: A person's security zone defined on the ground plane. It is parameterized by r, R, and ω, the conspicuous black dot representing the person's position. The arrow depicts the travel direction of the individual inferred from the tracking output.

The Nearness Diagram (ND) navigation is a navigation algorithm which is based on identification of free space areas and obstacles proximity based on some diagrams to define a set of situation which trigger specific motion laws. A situation is identified by the pose of the robot, the obstacle distribution, and the goal location. All known situations are used to build a decision tree beforehand. Then for any situation encountered, the tree is traversed based on binary decision rules that evaluate the current situation resulting (identifying) an associated action (control law) to use for this scenario. This specific obstacle avoidance has been chosen owing to its simplicity, real-time performance, and as it has been demonstrated to be an effective navigation method capable of avoiding collision in troublesome scenarios [57]. In our case, the only modification is: instead of seeing persons as just static obstacles, the algorithm will treat them as obstacles with special zone needs that depend on their motion direction.

All the discussed functionalities have been integrated in the presented platform. All implementations are modularized in a **G**eno**M** framework. For each block in figure 2, a dedicated **G**eno**M** module has been created. Each module is an independent, interacting module corresponding to a single functionality. This helps in standardizing all implementations in a solid software engineering framework. In due time, the implementations will be integrated in LAAS's opensource robotic software collection.

Table 7 summarizes where the major implementations are executing. The passers-by detection and tracking is made to run on the laptop so as to avoid massive data transfers (pertaining to the images) across the wireless connection because of a network's inherent delay. The complete system executes at approzimately 1 fps. The HOG based person detection running on the laptop, takes about 700 ms for two 640x480 images (camera 1 and 2) and is a major bottleneck.

To verify the seamless integration between the passers-by tracking and obstacle avoidance based on the presented security zone, the robot is simply

Table 7: Implementation Deployment

Functionality	Running-on
LRF leg detection	Robot PC
User (target) tracking	Robot PC
HOG person detection	Laptop
Foreground segmentation	Laptop
Multi-person Tracking	Laptop
Servoing actions	Robot PC

made to navigate from a starting point 'A' to an end point 'B', subject to interference from two passers-by. In repeated experiments of this sort, the robot successfully navigated from the start point to the end point adjusting its path to avoid the passers-by as anticipated. Figure 11 shows snapshot taken from the tracking module showing the position of the robot, the trajectory of the passers-by, the start and end point of the robot motion. The sequences clearly show how the robot rotated to go behind the passer-by tracked in red color [2].

Finally, the complete system, i.e. user following with passers-by avoidance, based on the passers-by tracking and defined security zone, is tested by making the robot follow the user amongst a total crowd of five people. In the experiment, Rackham was able to follow the person as anticipated without interfering with the motion of the passers-by. Snapshots taken from the live video captured during a run are shown in figure 12[2]. These results are satisfactory because the robot follows the user even in the event of disturbance by passers-by. But, the reaction time of the robot is still a little slow. In the current implementation this is mainly due to the computational cost incurred by the HOG based person detection which has a low frame rate and network latency. In its current form, this system could be extended for an area covered by similar number of cameras without problem. But, the system will have a problem if there be need to increase the number of cameras greatly for we have adopted a centralized fusion approach. At each cycle, information has to be sent to the central manager which will create a bottleneck when done over a network connection for large number of cameras. Hence, in this case a decentralize approach with decentralized processing at each camera could be favored. In summary, the presented system is well adapted for moderate

[2]The reader is referred to the URL homepages.laas.fr/aherbulo/videos/cviu/ for complete runs.

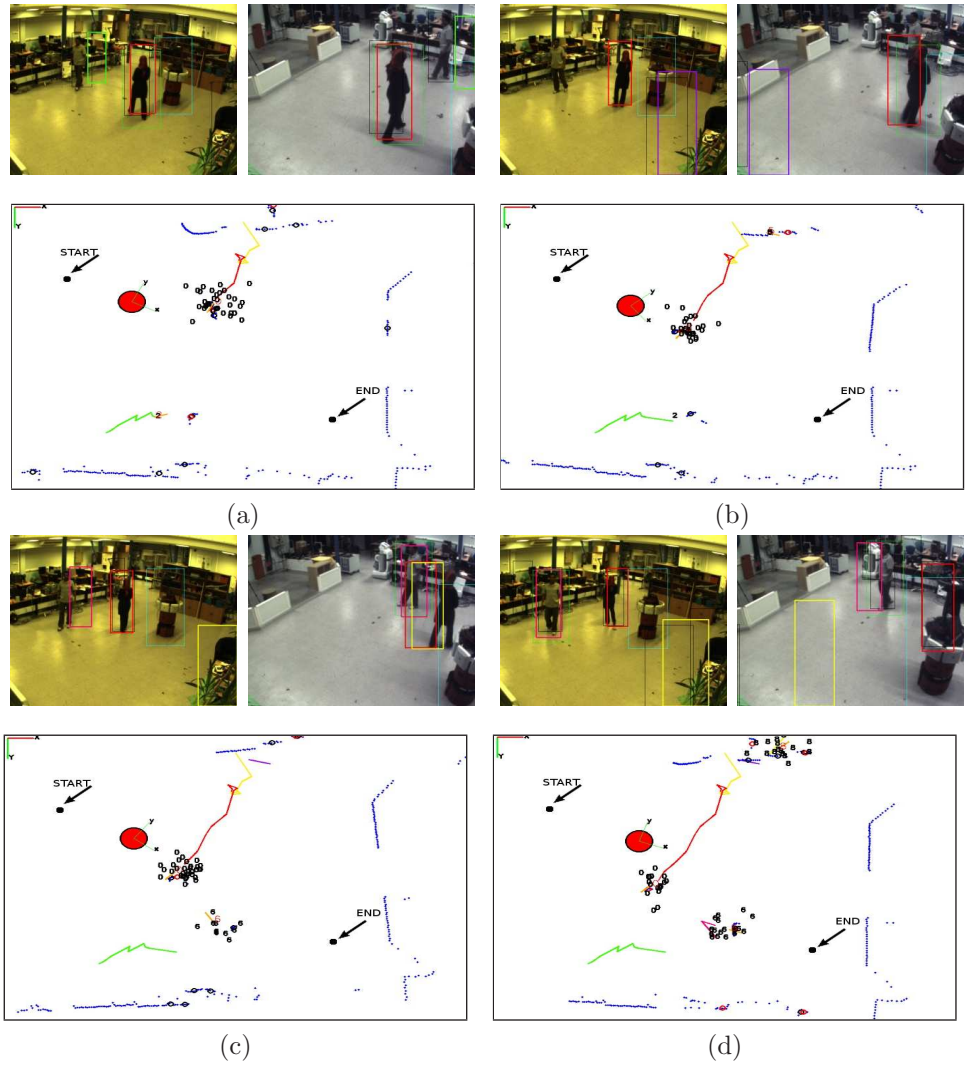
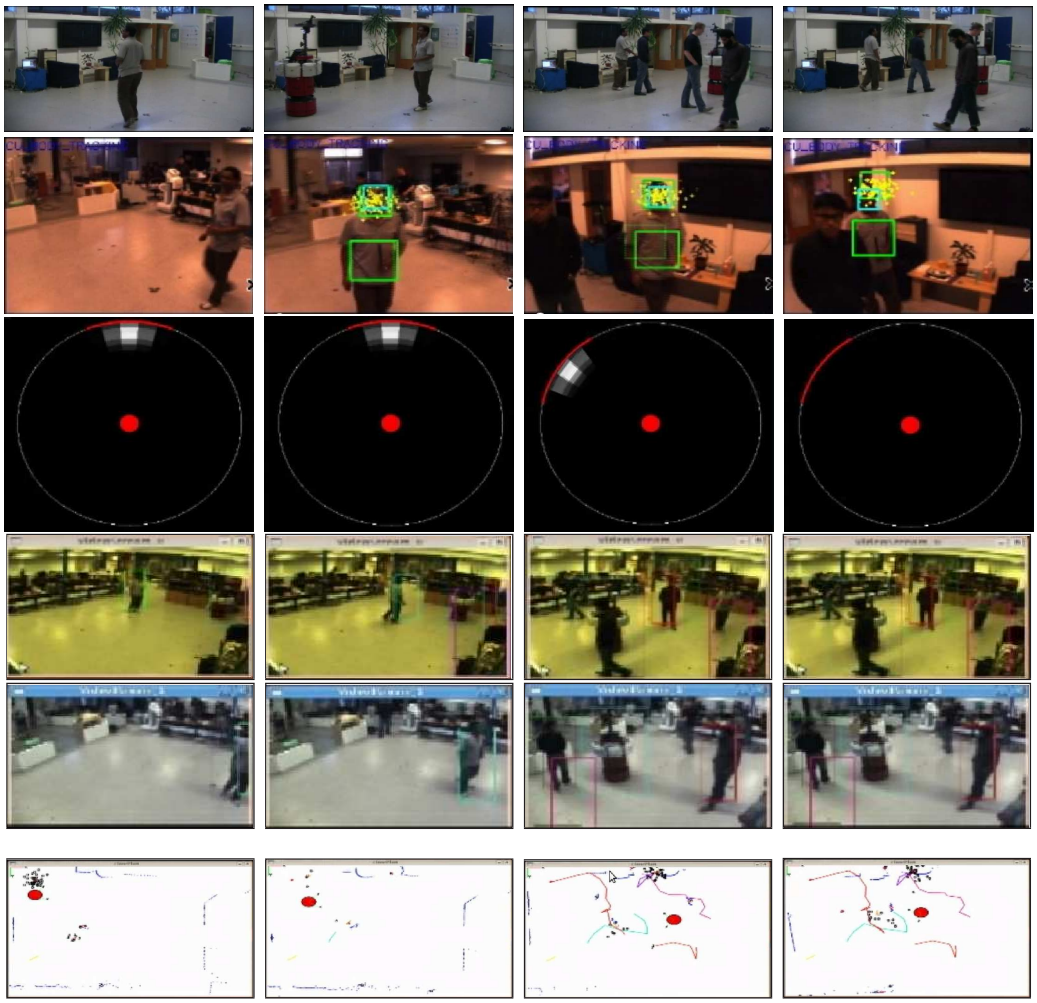


Figure 11: A sequence showing how the robot adjusted its trajectory to go to the end point without interfering with the passers-by. A security zone with $r = 0.5m$, $R = 1.5m$, and $\omega = 30^0$ is used in these runs.

size areas but not large halls like an Airport.

8. Conclusion and Perspectives

Person tracking provides important capabilities for human assistance in utilitarian populated areas. The work presented herewith makes its main



(a) 4.7*sec* (b) 10.8*sec* (c) 39.7*sec* (d) 40.6*sec*

Figure 12: Snapshot taken from the person following whilst passers-by avoidance live run. Each row corresponds to the feed from: external camera capturing the H/R situation (for demonstration purposes), the on-board vision, the RFID detection, the two deported (wall mounted) cameras, and ground plane showing tracking trajectories and laser scan, respectively. The time the shots were taken is shown in the caption. In (a), the user (target) person is seen entering the scene, with the RF detection indicating the direction of the tag held by the user. Tracking of the user can be seen in (b) with the yellow particles in the mobile camera feed, and a green bounding box in the wall mounted cameras feed. The robot actually starts to follow the user in a straight path as there are no passers-by in the vicinity. Passers-by start appearing in the workspace in (c) and their tracked trajectories can be seen on the ground plane showing the raw laser scan. (d) show tracking and avoidance instances during the experiment.

contribution in this vein by proposing a cooperative scheme between overhead cameras and sensors embedded on a mobile robot in order to track people in crowds. Our Bayesian data fusion framework outperforms (1) typical surveillance systems with only fixed cameras which can not handle dead spot, and (2) complete embedded systems without wide FOV and straightforward (re)-initialization ability. The presented off-line results are a clear indication of the framework's notable tracking performance. Our work extends the well-known intruder pursuit by a mobile robot to the tracking of all the passers-by in its vicinity and the disambiguation with its human user.

The proposed scheme has also been deployed on a development platform to verify its coherency and seamless integration amongst the different modalities. The live experiments demonstrate that the navigation task inherits the advantages of the various perceptual functions, thereby being able to robustly follow a given person whilst avoiding passers-by.

Near future investigations will focus on quantitative evaluation of the on-line (live) experiments in crowds. The HOG person detector will be replaced by a less time consuming detector to decrease the time processing and so further increase/improve the robot's reactivity. Further investigations will also concern the simultaneous interaction with multiple RF-tagged individuals. External PTZ cameras will also be added to (1) to increase the coverage range, and (2) achieve optimal camera assignment with respect to simultaneous and multiple observation goals during navigation.

References

[1] M. Takahashi, T. Suzuki, H. Shitamoto, T. Moriguchi, K. Yoshida, Developing a mobile robot for transport applications in the hospital domain, Robotics and Autonomous Systems 58 (7) (2010) 889 – 899.

[2] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, N. Hagita, A communication robot in a shopping mall, IEEE Transactions on Robotics 26 (5) (2010) 897 –913.

[3] A. Clodic, S. Fleury, R. Alami, R. Chatila, G. Bailly, L. Bréthes, M. Cottret, P. Danès, X. Dollat, F. Elise, I. Ferrané, M. Herrb, G. Infantes, C. Lemaire, F. Lerasle, J. Manhes, P. Marcoul, P. Menezes, V. Montreuil, Rackham: An interactive robot-guide, in: International Conference on Robot-Machine Interaction (RoMan'06), Hatfield, UK, 2006, pp. 502–509.

[4] W. Meeussen, E. Marder-Eppstein, K. Watts, B. P. Gerkey, Long term autonomy in office environments, in: ICRA 2011 Workshop on Long-term Autonomy, Shanghai, China, 2011, pp. 1–6.

[5] T. Germa, F. Lerasle, N. Ouadah, V. Cadenat, Vision and RFID data fusion for tracking people in crowds by a mobile robot, Computer Vision and Image Understanding 114 (6) (2010) 641–651.

[6] W. Choi, C. Pantofaru, S. Savarese, A general framework for tracking multiple people from a moving camera, IEEE Transactions on Pattern Analysis and Machine Intelligence (2012) 1doi:10.1109/TPAMI.2012.248.

[7] L. Spinello, K. O. Arras, People detection in RGB-D data, in: International Conference on Intelligent Robots and Systems (IROS'11), San Francisco, CA, USA, 2011, pp. 3838–3843.

[8] B. Khaleghi, A. Khamis, F. O. Karray, Multisensor Data Fusion: a Review of the State-of-the-Art, Information Fusion (2012),doi:10.1016/j.inffus.2011.08.001.

[9] C. Chia, W. Chan, S. Chien, Cooperative surveillance system with fixed camera object localization and mobile robot target tracking, in: T. Wada, F. Huang, S. Lin (Eds.), Advances in Image and Video Technology, Vol. 5414, Springer Berlin / Heidelberg, 2009, pp. 886–897.

[10] P. Chakravarty, R. Jarvis, External cameras and a mobile robot: A collaborative surveillance system, in: Australasian Conference on Robotics and Automation (ACRA'09), Sydney, Australia, 2009, pp. 1–10.

[11] W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviors, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 34 (3) (2004) 334 –352.

[12] X. Wang, Intelligent multi-camera video surveillance: A review, Pattern Recognition Letters 34 (1) (2013) 3 – 19. doi:10.1016/j.patrec.2012.07.005.

[13] B. Meden, F. Lerasle, P. Sayd, MCMC supervision for people reidentification in nonoverlapping cameras, in: Proceedings of the British

Machine Vision Conference (BMVC'12), BMVA Press, Surrey, UK, 2012, pp. 1–11.

[14] D. Arsic, E. Hristov, N. Lehment, B. Hornler, B. Schuller, G. Rigoll, Applying multi layer homography for multi camera person tracking, in: International Conference on Distributed Smart Cameras (ICDSC'08), Stanford, CA, USA, 2008, pp. 1–9.

[15] J. Cui, H. Zha, H. Zhao, R. Shibasaki, Tracking multiple people using laser and vision, in: International Conference on Intelligent Robots and Systems (IROS'05), Edmonton, Canada, 2005, pp. 2116–2121.

[16] D. Di Paola, A. Milella, G. Cicirelli, A. Distante, An Autonomous Mobile Robotic System For Surveillance Of Indoor Environments, International Journal of Advanced Robotic Systems 7 (1) (2010) 19–26.

[17] P. Cory, H. R. Everett, T. A. Heath-Pastore, Radar-based intruder detection for a robotic security system, in: H. M. Choset, D. W. Gage, P. Kachroo, M. A. Kourjanski, M. J. de Vries (Eds.), Mobile Robots XIII and Intelligent Transportation Systems, Vol. 3525, SPIE, 1999, pp. 62–72.

[18] L. Tmsuk Co., Security robot "t-34" from tmsuk co., ltd. and alacom co., ltd. (accessed: March 02, 2012 2004).
URL `http://www.tmsuk.co.jp/english/index.html`

[19] K. Arras, B. Lau, S. Grzonka, M. Luber, O. Mozos, D. Meyer-Delius, W. Burgard, Range-based people detection and tracking for socially enabled service robots, in: E. Prassler, M. Zllner, R. Bischoff, W. Burgard, R. Haschke, M. Hgele, G. Lawitzky, B. Nebel, P. Plger, U. Reiser (Eds.), Towards Service Robots for Everyday Environments, Vol. 76 of Springer Tracts in Advanced Robotics, Springer Berlin Heidelberg, 2012, pp. 235–280.

[20] J. Lee, T. Tsubouchi, K. Yamamoto, S. Egawa, People Tracking Using a Robot in Motion with Laser Range Finder, in: International Conference on Intelligent Robots and Systems (IROS'06), Beijing, China, 2006, pp. 2936–2942.

[21] A. Carballo, A. Ohya, S. Yuta, People detection using double layered multiple laser range finders by a companion robot, in: H. Hahn, H. Ko,

S. Lee (Eds.), Multisensor Fusion and Integration for Intelligent Systems, Vol. 35 of Lecture Notes in Electrical Engineering, Springer Berlin / Heidelberg, 2009, pp. 315–331.

[22] W. Zajdel, Z. Zivkovic, B. J. A. Kröse, Keeping track of humans: Have I seen this person before?, in: International Conference in Robotics and Automation (ICRA'05), Barcelona, Spain, 2005, pp. 2081–2086.

[23] P. Chakravarty, R. Jarvis, Panoramic vision and laser range finder fusion for multiple person tracking, in: International Conference on Intelligent Robots and Systems (IROS'06), Beijing, China, 2006, pp. 2949–2954.

[24] W. Choi, C. Pantofaru, S. Savarese, Detecting and tracking people using an rgb-d camera via multiple detector fusion, in: Workshop on Challenges and Opportunities in Robot Perception, at the International Conference on Computer Vision (ICCV), 2011, pp. 1076–1083.

[25] A. A. Mekonnen, F. Lerasle, I. Zuriarrain, Multi-modal people detection and tracking from a mobile robot in crowded environment, in: International Conference on Computer Vision Theory and Applications (VISAPP'11), Algarve, Portugal, 2011, pp. 511–520.

[26] C. Martin, E. Schaffernicht, A. Scheidi, H. Gross, Multi-modal sensor fusion using a probabilistic aggregation scheme for people detection and tracking, Robotics and Autonomous Systems 54 (9) (2006) 721–728.

[27] D. Beymer, K. Konolige, Tracking people from a mobile platform, in: International Symposium on Experimental Robotics (ISER'02), Sant'Angelo d'Ischia, Italy, 2002, pp. 234–244.

[28] M. Kobilarov, G. Sukhatme, J. Hyans, P. Bataria, People tracking and following with a mobile robot using an omnidirectional camera and a laser, in: International Conference on Robotics and Automation (ICRA'06), Orlando, Florida, USA, 2006, pp. 557–562.

[29] Z. Zivkovic, B. Kröse, Part based people detection using 2d range data and images, in: International Conference on Intelligent Robots and Systems (IROS'07), San Diego, CA, USA, 2007, pp. 214–219.

[30] J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Plötz, G. A. Fink, G. Sagerer, Multi-modal anchoring for human-robot interaction, Robotics and Autonomous Systems 43 (2-3) (2003) 133–147.

[31] N. Bellotto, H. Hu, Multisensor-based human detection and tracking for mobile service robots, IEEE Transactions on Systems, Man, and Cybernetics – Part B 39 (1) (2009) 167–181.

[32] Z. Zivkovic, B. Kröse, People detection using multiple sensors on a mobile robot, in: D. Kragic, V. Kyrki (Eds.), Unifying Perspectives in Computational and Robot Vision, Vol. 8, Springer US, 2008, pp. 25–39.

[33] K. Nakadai, K.-i. Hidai, H. Mizoguchi, H. G. Okuno, H. Kitano, Real-time auditory and visual multiple-object tracking for humanoids, in: International Joint Conference on Artificial Intelligence (IJCAI'01) - Volume 2, Seattle, WA, USA, 2001, pp. 1425–1432.

[34] X. Wu, H. Gong, P. Chen, Z. Zhi, Y. Xu, Intelligent household surveillance robot, in: International Conference on Robotics and Biomimetics (ROBIO'09), Guilin, Guangxi, China, 2009, pp. 1734–1739.

[35] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A. Fink, G. Sagerer, Audio-visual person tracking with a mobile robot, in: International Conference on Intelligent Autonomous Systems (IAS'04), Amsterdam, Netherlands, 2004, pp. 898–906.

[36] A. Treptow, G. Cielniak, T. Duckett, Real-time people tracking for mobile robots using thermal vision, Robotics and Autonomous Systems 54 (9) (2006) 729–739.

[37] M. Correa, G. Hermosilla, R. Verschae, J. Ruiz-del Solar, Human detection and identification by robots using thermal and visual information in domestic environments, Journal of Intelligent and Robotic Systems 66 (2012) 223–243.

[38] Microsoft corp. kinect for xbox.
URL `http://www.xbox.com/en-US/Kinect last accessed: March 02, 2012`

[39] M. Luber, L. Spinello, K. O. Arras, People tracking in rgb-d data with on-line boosted target models, in: International Conference on Intelligent Robots and Systems (IROS'11), San Francisco, CA, USA, 2011, pp. 3844–3849.

[40] Y. Y. Li, L. Parker, Detecting and monitoring time-related abnormal events using a wireless sensor network and mobile robot, in: International Conference on Intelligent Robots and Systems (IROS'08), Nice, France, 2008, pp. 3292–3298.

[41] F. Hoeller, D. Schulz, M. Moors, F. E. Schneider, Accompanying persons with a mobile robot using motion prediction and probabilistic roadmaps, in: International Conference on Intelligent Robots and Systems (IROS'07), San Diego, CA, USA, 2007, pp. 1260–1265.

[42] R. Alami, R. Chatila, S. Fleury, M. Ghallab, F. Ingrand, An architecture for autonomy, International Journal of Robotics Research 17 (1998) 315–337.

[43] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: International Conference on Computer Vision and Pattern Recognition (CVPR'01), Kauai, Hawaii, USA, 2001, pp. 511–518.

[44] J. Xavier, M. Pacheco, D. Castro, A. Ruano, Fast line, arc/circle and leg detection from laser scan data in a player driver, in: International Conference on Robotics and Automation (ICRA'05), Barcelona, Spain, April, 2005, pp. 3930–3935.

[45] P. Pérez, C. Hue, J. Vermaak, M. Gangnet, Color-based probabilistic tracking, in: European Conference on Computer Vision (ECCV'02), Copenhagen, Denmark, 2002, pp. 661–675.

[46] A. Manzanera, Sigma - delta background subtraction and the zipf law, in: Iberoamericann Congress on Pattern Recognition (CIARP'07), Valparaiso, Chile, 2007, pp. 42–51.

[47] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: International Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 886–893.

[48] D. Reid, An algorithm for tracking multiple targets, IEEE Transactions on Automatic Control 24 (6) (1979) 843–854.

[49] C. Rasmussen, G. D. Hager, Probabilistic data association methods for tracking complex visual objects, IEEE Transactions in Pattern Analysis and Machine Intelligence 23 (6) (2001) 560–576.

[50] M. Isard, J. MacCormick, Bramble: a bayesian multiple-blob tracker, in: International Conference on Computer Vision (ICCV'01), Vancouver, Canada, 2001, pp. 34–41.

[51] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, L. Van Gool, Robust tracking-by-detection using a detector confidence particle filter, in: International Conference on Computer Vision (ICCV'09), Kyoto, Japan, 2009, pp. 1515–1522.

[52] Z. Khan, T. Balch, T. Dellaert, Mcmc-based particle filtering for tracking a variable number of interacting targets, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (11) (2005) 1805–1918.

[53] F. Bardet, T. Chateau, D. Ramadasan, Illumination aware mcmc particle filter for long-term outdoor multi-object simultaneous tracking and classification, in: IEEE International Conference on Computer Vision, Kyoto, Japan, 2009, pp. 1623 –1630.

[54] K. C. Smith, Bayesian methods for visual multi-object tracking with applications to human activity recognition, Ph.D. thesis, Infoscience—Ecole Polytechnique Federale de Lausanne (Switzerland) (2007). URL `http://infoscience.epfl.ch/record/146296`

[55] K. Smith, D. Gatica-Perez, J. M. Odobez, Using particles to track varying numbers of interacting people, in: International Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 962–969.

[56] K. Bernardin, R. Stiefelhagen, Evaluating multiple object tracking performance: the CLEAR MOT metrics, EURASIP Journal on Image and Video Processing 2008 (2008) 1:1–1:10.

[57] J. Minguez, L. Montano, Nearness diagram (ND) navigation: collision avoidance in troublesome scenarios, IEEE Transactions on Robotics and Automation 20 (1) (2004) 45 – 59.