MULTI-MODAL PERSON DETECTION AND TRACKING FROM A MOBILE ROBOT IN A CROWDED ENVIRONMENT

A. A. Mekonnen[†‡], F. Lerasle[†,‡] and I. Zuriarrain[¶]
[†]*CNRS, LAAS, 7 Avenue du Colonel Roche, 31077 Toulouse Cedex 4, France*
[‡]*Université de Toulouse, UPS, INSA, INP, ISAE, LAAS, F-31077 Toulouse, France*
[¶]*University of Mondragon, Goi Eskola Politeknikoa, Mondragon, Spain*

Keywords: Multi-person tracking, Multi-modal data fusion, MCMC particle filtering, Interactive robotics.

Abstract: This paper addresses multi-modal person detection and tracking using a 2D SICK Laser Range Finder and a visual camera from a mobile robot in a crowded and cluttered environment. A sequential approach in which the laser data is segmented to filter human leg like structures to generate person hypothesis which are further refined by a state of the art parts based visual person detector for final detection, is proposed. Based on this detection routine, a Monte Carlo Markov Chain (MCMC) particle filtering strategy is utilized to track multiple persons around the robot. Integration of the implemented multi-modal person detector and tracker in our robotic platform and associated experiments are presented. Results obtained from all tests carried out have been clearly reported proving the multi-modal approach outperforms its single sensor counterparts taking detection, subsequent use, computation time, and precision into account. The work presented here will be used to define navigational control laws for passer-by avoidance during a service robot's person following activity.

1 INTRODUCTION

Currently, there is more demand to use robots in everyday life, a demand for their introduction into human all day environments. For this task, robots should be able to interact with humans at a higher level with more natural and effective interaction. One such interaction, the ability of a mobile robot to automatically follow a person in public areas, is a key issue to effectively interact with the surrounding world. Recently various researchers have reported successful person following activities from a mobile robot (Germa et al., 2009), (Calisi et al., 2007), (Chen and Birchfield, 2007). A key point in person following task is safe interaction as the workspace at any moment is shared by humans and the robot. The robot should be capable of avoiding all passers-by in the environment in a socially acceptable manner while carrying out the activity. Some authors addressed this as static obstacle avoidance considering people as static obstacles *e.g.*(Calisi et al., 2007). We argue otherwise, an effective collision avoidance not only has to circumvent static objects in the environment, but it also has to take the dynamics of the persons in the surrounding into account. This entails for perception of the whereabouts and dynamics of humans sharing the workspace. To the best of our knowledge, an assistant robot capable of following a given person taking the dynamics of the passers-by into consideration and avoiding them in a socially acceptable way does not yet exist (Fong et al., 2003).

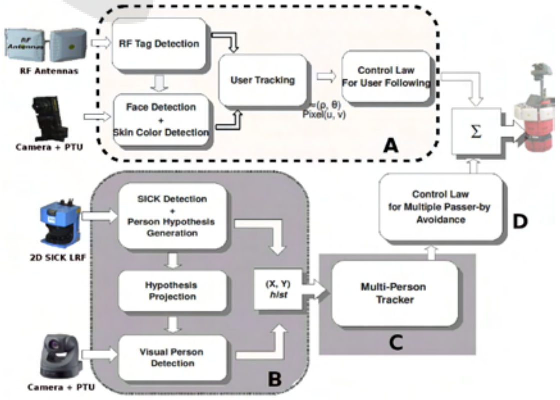


Figure 1: A block diagram of our complete envisaged system: Person Following with passer-by avoidance in a socially acceptable manner.

A block diagram of our complete envisaged system is shown in figure 1. The block diagram represents a person following activity with passer-by avoidance in a socially acceptable manner (keeping a social distance from surrounding persons while at the same time taking their dynamics into consideration)

A. Mekonnen A., Lerasle F. and Zuriarrain I..
MULTI-MODAL PERSON DETECTION AND TRACKING FROM A MOBILE ROBOT IN A CROWDED ENVIRONMENT.
DOI: 10.5220/0003367705110520
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2011), pages 511-520
ISBN: 978-989-8425-47-8

by a service robot. The person following activity, the area labeled ‘A’ in figure 1, has been successfully addressed in (Germa et al., 2009). The work presented here addresses detection and tracking of multiple people around the robot, the shaded areas: block ‘B’ and ‘C’ in figure 1, while the control law for passer-by avoidance will be presented in future works. In a nutshell, the objective of the work presented in this paper is detection and tracking of people in the robot vicinity maintaining a correct trajectory of all tracked people. It is aimed for defining control laws for socially acceptable passer-by avoidance, during person following activity, in a crowded public environment.

Automated person detection/tracking finds its applications in many areas including robotics, video-surveillance, pedestrian protection systems, automated image and video indexing. Contrary to video-surveillance applications where conventional background subtraction can be used, person detection is more challenging in mobile robotics due to sensor limitations, short fields of view and motion of embedded sensors, and computational requirements for reactive response acceptable by humans. All these challenges make successful person detections and tracking based on a single sensor very difficult. Several works on person detection in the robotic community are based on vision and Laser Range Finders (LRFs)(Schiele et al., 2009). Eventhough vision based person detection yields a lot of information, detections are very sensitive to illumination variation, deformations, and partial occlusions on top of the associated high computational cost. Person detection based on LRFs are computationally cheap and insensitive to illumination. But their information content is not discriminative enough for robust detection unless used in a non-cluttered environment with a priori learnt environment map which is not realistic for crowded dynamic scenes. For real world scenarios, well established approaches combine inputs from more than one sensory channel, a majority of the works combining vision and Laser, *e.g.* (Zivkovic and Kröse, 2007) (Spinello et al., 2008). In this vein, we propose a multi-modal person detector that uses a 2D SICK Laser Range Finder (LRF) and a visual camera for detecting multiple persons around the robot. A sequential approach in which the laser data is segmented to filter human leg like structures to generate person hypothesis which are further refined by a state-of-the-art parts based visual person detector for final detection, is proposed. To be able to make spatio-temporal analysis of the targets, we have also employed tracking based on the detections.

The literature in multi-target tracking contains different approaches, most commonly: Multiple Hypothesis Tracker (MHT)(Reid, D., 1979), Joint Probabilistic Data Association Filter (JPDAF)(Rasmussen and Hager, 2001), centralized (Isard and MacCormick, 2001) and decentralized particle filters (PFs) (Breitenstein et al., 2009), and MCMC PF (Khan et al., 2005). 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, RJMCMC PF, an extension of MCMC to variable dimensional state space, has been pioneered to perform successful tracking (Khan et al., 2005). The MCMC PF frame work including RJMCMC PF has been validated in video surveillance context solely on visual data, *e.g.*(Smith et al., 2005). Inspired by this, we have used RJMCMC PF for multi-person tracking driven by our multi-modal detector with sensors embedded on a robot. Implementation details along with integration in our robotic platform, associated experiments, and evaluation results are presented, proving the proposed approach outperforms its single sensor counterparts taking detection, subsequent use, computation time, and precision into account.

This paper is structured as follows: section 2 discusses our multi-modal person detector implementation while section 3 presents our implementation of the RJMCMC PF tracker. Integration of the developed functionalities in our robotic platform, associated experiments, results, and discussions are presented in section 4. Finally, section 5 summarizes the presented work and highlights possible future investigations.

2 MULTI-MODAL PERSON DETECTOR

Our multi-modal person detector is based on a 2D SICK Laser Range Finder and a visual camera.

2.1 SICK-based Detector

Recently, 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. Detection of a person from LRF information hence proceeds by trying to detect shapes of a person in the scan data at the height the scan is performed. In the context of this work, leg detection will be considered as the laser scanner used is positioned at a height of 38 cm above the ground.

Our robotic platform, Rackham (presented in §4.1), has a SICK LMS200 2D laser range finder that swipes an arc of 180^o measuring the radial distance of obstacles in a set angular resolution of 0.5^o. The detection makes use of geometric properties of leg scans highlighted in (Xavier et al., 2005) with no a priori environment map assumption. Though if a 2D map of the environment, made of line segments, is available, all points not lying on the map are filtered to be considered further. The detection proceeds in three steps:

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.

Blob Filtering. The blobs formed are filtered using geometric properties outlined in (Xavier et al., 2005). 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 (Xavier et al., 2005).

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.

This detection system has some drawbacks, namely: false detection of table legs, chair legs, and other narrow objects with circular pattern. People standing with closed legs or wearing long skirts do not yield appropriate leg signatures needed by the detector, so are classified as negative instances resulting in false-negatives. On top of these, it is not possible to know which leg detections correspond to which person in the presence of multiple people, making associations of each legs in consecutive frames difficult. This mode of detection is different from the combined detector presented in §2.3 in that it makes use of all the geometric properties strictly for leg detection.

2.2 Visual Detector

Recently, remarkable advances have been made in automated visual person detection, (Dalal and Triggs, 2005), (Laptev, 2006), and recently (Felzenszwalb et al., 2010). For visual person detection, we have used our complete C implementation of the state-of-the-art person detector, Felzenszwalb's person detector with discriminatively trained part based models. The detector is based on mixtures of multi-scale deformable parts models that have the ability to represent a highly variable object class like that of a person. The resulting person detector is efficient, accurate, and has achieved state-of-the-art results in the PASCAL VOC competition and the INRIA person dataset[1]. Briefly speaking, the detector uses contrast sensitive and insensitive Histograms of Orientation Gradients (HOGs) with analytically reduced dimension as features. A person is modelled using a star-structured part based model defined by a root filter and a set of parts filters with associated deformation models. Compared to full body detection approaches, (Dalal and Triggs, 2005), (Laptev, 2006), this body parts based detector is more robust to partial occlusions. The person model currently implemented consists of mixtures of two models each of which have one coarse root filter that approximately covers an entire person and six high resolution parts filters that cover smaller parts of the object. For details on this person detector, the reader is referred to (Felzenszwalb et al., 2010).

In this work a person model trained with the Pascal VOC 2008 dataset and provided with the Matlab open source (Felzenszwalb et al., 2009) is used. Unfortunately, the C implementation of the person detector takes about 4.6 seconds to detect persons on a 320x240 image on a PIII 850 MHz computer with 6 levels in each octave of the feature pyramid. This computation time is not acceptable for the task at hand, navigation in a crowded environment, and entails further improvements to speed the detection process.

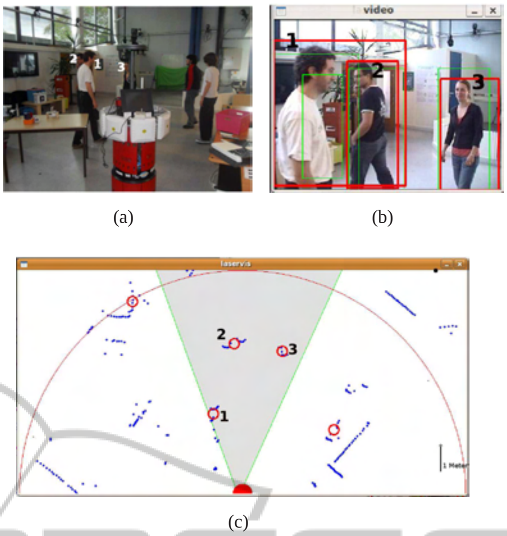
2.3 Combined Detector

The person detection from LRF suffer from false positives due to structures resembling that of a person leg, mis-detections due to closed legs or long skirt, and do not carry enough information to discriminate detections between multiple persons. On the other hand, the visual person detector (Felzenszwalb et al., 2010) is not readily applicable for the objective at hand due to computation time requirement. To make use of the

[1] See the URL http://pascal.inrialpes.fr/data/human/

two detectors in a complementary fashion, a multi-modal detector is implemented. The block labelled 'B' in figure 1 shows a block diagram of the overall multi-modal detector. Similar to (Cui et al., 2005) and (Spinello et al., 2008), the proposed approach is to define region of interests, henceforth referred as person hypothesis, using the detections from the laser scanner and then validate this by using the visual person detector on these regions. For this, first the geometric criteria to detect persons from the 2D laser scanner within the camera field of view region are relaxed to have a 100% person detection while at the same time having many false positives. For every hypothesis, a virtual rectangle conforming to an average person height of 1.8 m with an aspect ratio of 4:11 (width:height) is positioned at the precise distance obtained from the laser assuming a flat world. Then each virtual rectangle is projected on to the image, thanks to a complete calibrated camera system, defining a rectangular search region on the image. The parts based visual person detector is used to evaluate these defined regions. All the regions confirmed to contain persons are labelled as detections while those hypothesis not confirmed by the visual detector are discarded as false alarms. The main advantage of using the defined region of interests is the reduced computation time. Neither all the levels of the feature pyramid nor model scores at all possible positions on the feature pyramid need be computed. In the region outside the camera field of view, detection is solely based on the laser range finder as described in subsection 2.1. Note that this mode of operation differs from the Laser only based detection, explained in subsection 2.1, in the region within the camera field of view as loose geometric constraints are made use of for speeding the visual person detector.

Finally, the multi-modal person detector provides a list of detected targets with their precise locations, (x, y), in the ground plane with respect to the robot, and a normalized histogram of the image patch (if the detection occurred within the field of view of the camera) to the Multi-Person Tracker. Figure 2 shows a typical instance of the multi-modal detector. Figure 2(a), shows the Human-Robot situation, 2(b) detected persons by the multi-modal detector with bounding boxes projected on the image plane, and 2(c) shows the raw laser data (in blue) with the corresponding person detections (in red) in the ground plane. The shaded area in figure 2(c) is the camera field of view whereas the robot is depicted as the red object in the center of the arc.



(a) (b)

(c)

Figure 2: An instance of the multi-modal person detector with Human-Robot situation.

3 MULTI-PERSON TRACKER

The Multi-Person Tracker (MPT) 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 in the vicinity of the robot and within the field of view of the utilized sensors based on the detector outputs.

3.1 Formalism

In object tracking in general, the primary goal is to determine the posterior distribution $P(X_t|Z_{1:t})$ of a target state X_t at the current time t, given the observations sequence $Z_{1:t} = \{Z_1, Z_2, ..., Z_t\}$. Under the Markovian target motion assumption, the Bayes filter offers a concise way to express the tracking problem. Particle Filters offer approximations of the Bayes filter by propagating N number of particles over time to approximate the posterior $P(X_t|Z_{1:t})$ as a sum of Dirac functions, such that: $P(X_t|Z_{1:t}) \approx \frac{1}{N}\sum_{n=1}^{N}\delta(X_t - X_t^n)$ where X_t^n denote the n^{th} particle. In multi-target tracking, the state encodes the configuration of the 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. In MCMCPF, the inefficient importance sampling of a classical Particle Filters is replaced with a more efficient MCMC sampling step. MCMC methods define a Markov Chain over the space configuration X_t^n such that the stationary distribution of the chain is equal to the desired posterior.

Reversible Jump Monte Carlo Markov Chain (RJMCMC) PF is an extension of MCMC PF that accounts for the variability of the tracked targets by defining a variable dimension state space. In this case, the state space dimension is considered as a union of several subspaces. Whenever a new person enters the scene, the state "jumps" to a larger dimension subspace and there will be a "jump" to a lower dimension subspace whenever a tracked person leaves the scene. An important point in RJMCMC is the reversibility of the proposals that vary the dimensionality of the state space exploration. Any jump between subspaces must have a corresponding reverse jump to prevent the search chain from getting stuck in local minimum. These moves that guide the state space exploration are referred as proposal moves. A common technique that simplifies both the transition of the new proposed state hypothesis X^* from X and evaluation of the acceptance ratio is, for the state transition model to consider only changes to a randomly chosen subset of the state (in the case of multi-target tracking, this translates into changing a single target per iteration). In cases where interaction between different targets is likely to occur, an Interaction Model should be included to maintain tracked target identity.

Algorithm 1: RJMCMC Particle Filter.

```
Input: Particle set at time $t-1$ : $\{X_{t-1}^n\}_{n=1}^N$
    Prediction: generate a prediction set at time $t$ : $\{X_{t-1}^{n*}\}_{n=1}^N$ according to the system dynamics $Q(X_t^n|X_{t-1}^n)$.
    Init: $X_t^0 = X_t^{r*}, r \in \{1,...,N\}$
for $i = 0$ to $N + N_B$ do
    ▷ Choose a move $m \in \{add, update, remove\} \sim q_m$.
    if $m ==$ 'add' then
        ▷ $X^* = \{X_t^{i-1}, x_{I^{i-1}+1}\}$, with the new target $x_p = x_{I^{i-1}+1}$ and $I^{i-1}$ representing the number of persons hypothesized by $X_t^{i-1}$
    else if $m ==$ 'remove' then
        ▷ $X^* = \{X_t^{i-1} \setminus x_p\}$ where $p \in \{1,...,I^{i-1}\}$
    else if $m ==$ 'update' then
        ▷ Randomly choose a tracked person from $X_t^{i-1}$.
        ▷ Replace the person's state in $X_t^{i-1}$ with a randomly chosen state corresponding to this person in the prediction set $\{X_{t-1}^{n*}\}_{n=1}^N$, proposing $X^*$.
    end if
    ▷ Compute Acceptance Ratio :
        $\beta = min(1, \frac{\pi(X^*)Q_{ind*}(X_t^{i-1}|X^*)\Psi(X^*)}{\pi(X_t^{i-1})Q_{ind}(X^*|X_t^{i-1})\Psi(X_t^{i-1})})$
    where $ind \in \{add, update, remove\}$ and $ind^*$ denotes the reverse operation.
    if $\beta \geq 1$ then
        ▷ $X_t^i = X^*$
    else
        ▷ Accept $X_t^i = X^*$ with probability $\beta$ or reject and set $X_t^i = X_t^{i-1}$
    end if
end for
▷ Discard the first $N_B$ samples of the chain (burn-in).
▷ Compute the MAP estimate, $\hat{X} = E_{p(X_t|Z_{1:t})}[X_t] = \arg\max_{X_t^i}[count(x_k^i)]$
Output: Particle Set at time $t$: ${X_t^n}_{n=N_B+1,...,N_B+N}$ and MAP estimate, $\hat{X}$.
```

3.2 Implementation

Our RJMCMC PF tracker is driven by the multi-modal detector described in §2. To handle the variability of the tracked targets three sets of proposal moves are utilized in the RJMCMC PF: $\{$*Add, Update, Remove*$\}$. A Markov Random Field is also used to model the interactions amongst targets. The complete principle of our tracker is presented in Algorithm 1. Roughly, the algorithm iterates $N + N_B$ times proposing new state based on the previous one. N is the number of particles whereas N_B represents the number of burn-in iterations needed to converge to stationary samples. Each subsection below gives an overview of part of the algorithm in detail.

3.2.1 State Space

The state vector of a single hypothesis n at a certain time t in our tracker is made of the joint state vectors of the tracked persons (encodes the entire configuration): $X_t^n = \{I_t^n, x_{t,i}^n, i \in 1,...,I_t^n\}$, where I_t^n is the number of tracked persons, N is the total number of hypotheses (particles), and x_t^n is the state vector of individual persons. Since our aim is to outline trajectories of persons around the robot, the tracking is done on the ground plane. Hence, the state vector of an individual person is represents as $(Id, \mathrm{x}, \mathrm{y})$ in the ground plane with respect to the robot. Formally, the i^{th} state vector of a single person in hypothesis n at time t is a 2D vector represented as: $x_{t,i}^n = \{Id_i, \mathrm{x}_{t,i}^n, \mathrm{y}_{t,i}^n\}$.

3.2.2 Proposal Moves

At each iteration of the RJMCMC PF, a proposal move on only one randomly chosen dimension is proposed. Recall that three sets of move are considered, namely: $m = \{Add, Update, Remove\}$. The choice of the proposal privileged in each iteration is determined by q_m, the jump move distribution. The probabilities of Add, $Update$, and *Remove* are set to 0.2, 0.6, and 0.2 respectively. These proposal moves make use of the proposal densities, $Q()$, associated with them. The proposal densities make use of two masking maps: a map made from detected targets, and a map made from the tracked (MAP estimate) targets. Assuming the number of detected persons at time t is M_t, each detection can be represented as $p = (x, y)$ to make an associated mask map as a Gaussian mixture with each detection as a Gaussian having mean p and assumed variance Σ (equation 1). Similarly, each tracked target in the MAP estimate at $t-1$ is used to make a masking map as a Gaussian mixture; $\hat{x}_j$ as mean values, where $j \in 1,..,N_t$ and N_t is total number of tracked targets (equation 2).

$$S_t^d(Z_t) = \sum_{j=0}^{M_t} \mathcal{N}(;p_j,\Sigma) \tag{1}$$

$$S_t^{map}(\hat{X}_{t-1}) = \sum_{j=0}^{N_t} \mathcal{N}(;\hat{x}_j,\Sigma) \tag{2}$$

Add. The add move, randomly select a detected person, x_p, from the multi-modal detector and appends its state vector on X_t^{i-1} resulting in a proposal state X^*. The proposal density driving the *Add* proposal, when computing the acceptance ratio $Q_{Add}(X^*|X_t^{i-1})$, is given in equation 3. The mask map from detected targets is multiplied by a map derived from the tracked targets mask map. This distribution will have higher values whenever an add is proposed on locations conforming to detected targets that are not yet being tracked.

$$Q_{add}(X^*|X_t^{i-1}) = S_t^d(Z_t) * (1 - S_t^{map}(\hat{X}_{t-1})) \tag{3}$$

Figure 3 illustrates the derivation of $Q_{add}(X^*|X_t^{i-1})$ at a certain time t. Figure 3(a) bottom shows an inverted mask derived from the tracked targets at time $t-1$ and figure 3(b) shows the detected targets. Finally, the distribution $Q_{add}(X^*|X_t^{i-1})$ is derived by multiplying both (figure 3(c) bottom). The derived distribution shows higher values in the region near the detected target that is not being tracked, favoring its addition. The top figure in 3(c) shows the effect of the mask on the actual video image and is presented here solely for clarity purposes.

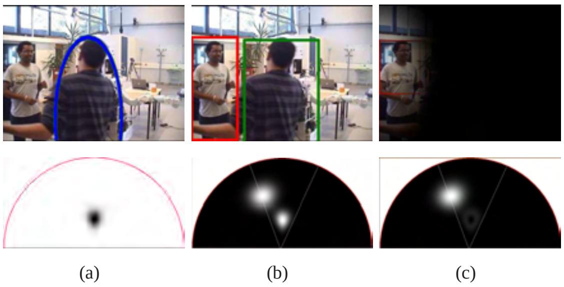

(a) (b) (c)

Figure 3: Derivation of $Q_{add}(X^*|X_t^{i-1})$ from tracked targets and detection. White intensity value represents high value whereas black is for low value.

Remove. The remove move, randomly selects a tracked person from the particle being considered, X_t^{i-1}, and removes it, proposing a news state X^*. Contrary to the add move, the proposal density used when computing the acceptance ratio, $Q_{Remove}(X^*|X_t^{i-1})$ (equation 4), is given by the mask map from the tracked targets multiplied by a map driven from the detected targets. This density assures targets that are not detected but are still being tracked have higher values. Figure 4 depicts derivation of the $Q_{Remove}(X^*|X_t^{i-1})$ distribution. A tracked target has just left the scene but the tracker still has the person in its state (figure 4(a) bottom). The detector returns one detection corresponding to the person still in the scene (figure 4(b)). As illustrated, the final $Q_{Remove}(X^*|X_t^{i-1})$, figure 4(c) bottom, shows high values for the target which left the scene favoring its removal. Figure 4(c) top illustrates the effect of the mask on the video feed, all black meaning no target in the camera field of view should be removed.

$$Q_{remove}(X^*|X_t^{i-1}) = (1 - S_t^d(Z_t)) * S_t^{map}(\hat{X}_{t-1}) \tag{4}$$

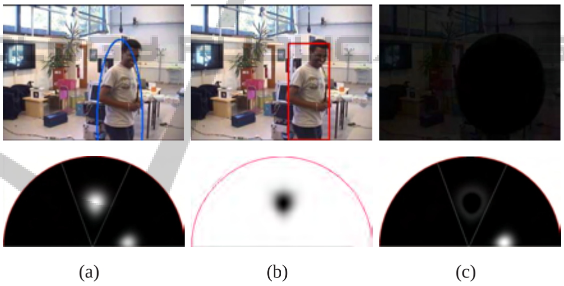

(a) (b) (c)

Figure 4: Illustration of $Q_{Remove}(X^*|X_t^{i-1})$ derivation. White intensity value represents high value whereas black is for low value.

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

3.2.3 Interaction Model

Similar to (Khan et al., 2005) and (Smith et al., 2005), a Markov Random Field (MRF) is adopted to address the interactions between nearby targets. The MRF is defined on an undirected graph, with targets defining the nodes of the graph, and links created at each time-step between pairs of proximate targets. A pairwise MRF where the cliques are restricted to the pairs of nodes that are directly connected to the graph, is implemented as part of our tracker. For a given state X, the MRF model is given by equation 5. $\phi(x_i, x_j)$ evaluates to zero if two targets are in the same position, penalizing fitting of two trackers to the same object during overlap (interaction).

$$\Psi(X) = \Pi_{i \neq j} \phi(x_i, x_j)$$
$$\phi(x_i, x_j) = 1 - exp(-(\frac{d(x_i, x_j)}{\sigma})^2) \tag{5}$$

where $d(x_i, x_j)$ is Euclidean distance, $i, j \in \{1, \ldots, N\}$, and N number of targets in X.

3.2.4 Observation Likelihood

The likelihood measure is derived from the 2D laser range raw data. Every segmented blob is filtered to keep blobs within a range of radius. This filters out laser data pertaining to walls, thin table or chair legs, and other wide structures. Then every filter blob is represented as a Gaussian centered on the centroid of the blob. The complete mixture of Gaussians makes up the likelihood map for our tracker. Given a state X, its likelihood is evaluated as the sum of likelihood values on the position of each target averaged over the number of targets (equation 6).

$$\pi(X^*) = \frac{1}{N_t} \sum_{i=0}^{N_t} s_t^{lik}(Z_t)|_{(x_i, y_i)}$$
$$s_t^{lik}(Z_t) = \sum_{j=0}^{N_b} \mathcal{N}(; Z_{t,j}, \Sigma) \tag{6}$$

Where N_t is the number of targets in X^*, and N_b is the number of blobs formed from the laser reading (Z_t).

4 EXPERIMENTS

4.1 Robotic Platform

The target robotic platform, Rackham, is an iRobot B21r mobile platform (figure 5). Its standard equipment has been extended with one pan-tilt Sony EVI-D70 camera, one digital camera mounted on a Directed Perception pan tilt unit(PTU), one ELO touch screen, a pair of loudspeakers, an optical fiber gyroscope, a Wireless Ethernet, and an RF system for detecting RFID tags. It integrates two PCs (one mono-CPU and one bi-CPUs PIII running at 850 MHz). Rackham also has an LMS200 SICK Laser Range Finder as its standard equipment. All these devices give Rackham the ability to operate in public areas as a service robot. The digital camera with the Directed Perception PTU is dedicated for the person following activity along with the RF system, whereas the Sony EVI-D70 camera is used for the multi-person (passer-by) detection and tracking. Rackham's software architecture is based on the GenoM architecture for autonomy (Alami et al., 1998). All its functionalities have been embedded in modules created by **GenoM** using C/C++ interface. Accordingly, the multi-modal person detector and MCMC PF tracker described are implemented as a **GenoM** modules.

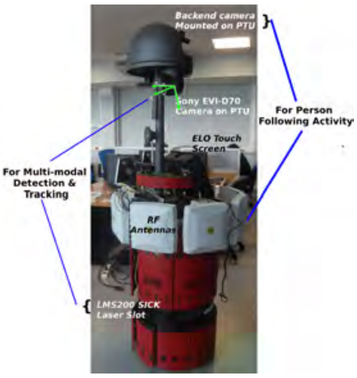


Figure 5: Rackham with its onboarded sensors.

4.2 Offline Evaluations

The offline evaluation corresponds to the evaluation of both the multi-modal person detector and the RJMCMC PF tracker offline using real data acquired with Rackham.

4.2.1 Multi-modal Detector Evaluations

In all the experiments, a 5 meter radius around the robot is considered for detection and tracking. The camera has a 45^o field of view, leaving the rest of laser scanner field of view, 135^o, for laser only detection. To evaluate the multi-modal person detector a dataset containing a total of 2872 frames is used. To quantify performance of the multi-modal person detector, two measures namely True Positive Rate (TPR) and False Positive Per Image (FPPI) are used.

- True Positive Rate (TPR): computes the ratio of correctly detected targets to the total number of targets present averaged over the entire dataset, i.e. $\frac{1}{J_t} \sum_{k,j} \delta_{k,j}$ where $\delta_{k,j} = 1$ if a target is detected in frame k or 0 otherwise. J_t is the total number of targets present in the entire dataset.
- False Positive Per Image (FPPI): computes the false positive occurrence per frame averaged over the entire dataset, *i.e.* $\frac{1}{K} \sum_{k,j} \delta_{k,j}$ where $\delta_{k,j} = 1$ if a target j is detected when there is actually no target in frame k or 0 for correct detection. K is the total number of frames in the entire dataset.

All the 2872 frames were hand labelled for (x, y) positions of persons on the ground plane based on the

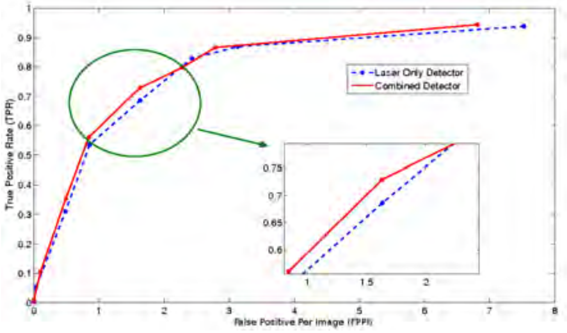
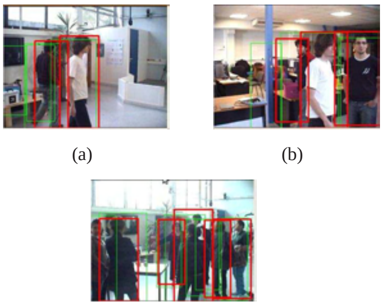


Figure 6: ROC curve (TPR Vs FPPI) comparing the performance of both LRF only detector, and the multi-modal detector.

laser data. A True Positive occurs whenever a detection is within 30 cm radius of the ground truth. A Receiver Operator Curve (ROC), TPR vs FPPI ROC graph shown in figure 6, is generated by relaxing and/or straining the geometric constraints for leg detection. To verify that the multi-modal detector is superior than the Laser only based detector, the experiment has also been done on the Laser only based detector. Hence, the ROC curve is generated for the LRF only based person detector (§2.1) and for the multi-modal person detector (§2.3).

Looking at the ROC curve in figure 6, it can be seen that, the addition of the visual detector improves the overall detection performance. On top of this performance improvement, a rich discriminative information is obtained whenever a target is within the field of view of the camera. Balancing True Detection with False Positive rate, the multi-modal person detector is set to operate at a point with *TPR = 0.72 and FPPI = 1.6*. Sample detections obtained operating the detector at this point are shown in figure 7.



(a)

(b)

(c)

Figure 7: Sample person detection with the multi-modal detector.

4.2.2 Multi-person Tracker Evaluations

Similarly, to evaluate the performance of the MCMCPF multi-person tracker, two complete sequences are used.

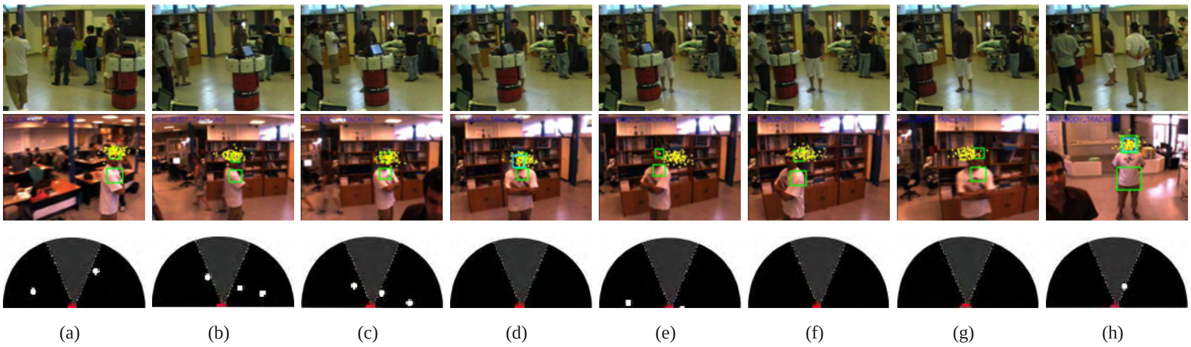
- Sequence I. A sequence of 785 frames containing two moving targets.
- Sequence II. A sequence of 507 frames with two moving targets but once in a while other targets appear and disappear in the tracking area.

As a performance measure, the following three measures are computed.

- 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 time t, else 0. $J_t = \sum_{k,j} j_k$, and j_k represents the number of persons in the tracking area at frame k.
- 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.
- Precision Error (PE): measures how precisely the targets are tracked, as the sum of the squared error between tracker position estimate and ground truth averaged over the entire sequence.

For each sequence, a hand labeled ground truth with (x,y) position and unique Id for each person is used. Similar to the detection, a person is considered to be correctly tracked (True Success), if the tracking position is within a 30 cm radius of the ground truth. All the Gaussians used to make associated distributions are constructed in polar form (ρ,θ) with standard deviation of $\sigma_\rho = 30cm$ and $\sigma_\theta = 0.157rad$. These values are set to account for a single walk (of an average person) uncertainty. Whenever a target is in the camera field of view, a histogram of the region subtended by the target is cached in memory. This histogram is used to overcome discontinuities in tracking when a tracked target is removed and initialized as a new target due to subsequent misdetections. It is also used to make association distinctions whenever targets come close by and some configuration within the state sample another target's state.

For evaluation, each sequence is run ten times to account for the stochastic nature of the filter. Results are reported as mean value and associated standard deviation in table 1. The results show that our multi-modal person tracker performs well on the two sequences used for evaluation, with a 73.3% True Detection on the first sequence. What should be highlighted here is that, the detector plus tracker makes no use of a priori knowledge of the environment. The environment the experiments were carried out is a highly cluttered environment containing many artifacts that resemble the leg of a person and the field of

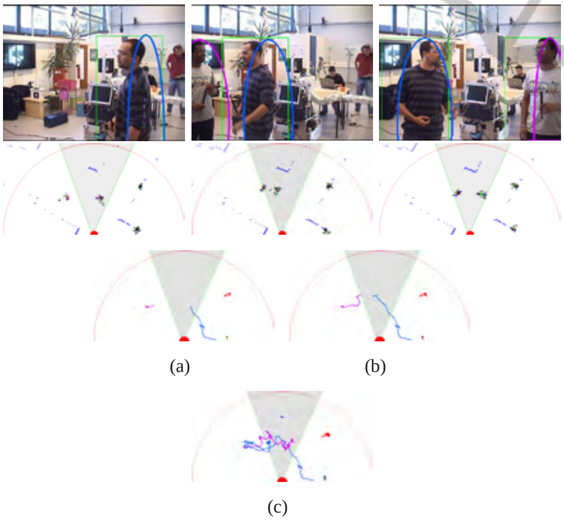


(a) (b) (c) (d) (e) (f) (g) (h)

Figure 9: A sequence of frames from online obstacle avoidance scenario based on the multi-modal person detector. The top row shows Human-Robot situation, the middle user tracking for person following, and the third multi-modal person detections on the ground plane.

Table 1: Results of the MCMCPF multi-tagged person tracker.

MCMCPF Person Tracking Results			
Seq.	TSR	GR	PE (cm)
I	0.733 ± 0.074	1.221 ± 0.078	7.93 ± 0.68
II	0.62 ± 0.078	1.355 ± 0.297	8.37 ± 1.13



(a) (b)

(c)

Figure 8: Sample snap-shots taken from the multi-person tracking on sequence I at the 35^{th}, 51^{th}, and 259^{th} frames respectively.

view of the camera is very narrow. The average position precision of the tracker is also less than $9cm$. An average Id switch per sequence of one for the first sequence and two for the second has also been observed.

Figure 8 shows sample snap-shots[2] taken from tracking runs of sequence 1. The top row shows the tracking on the video feed, the middle shows the particle swarm, and the bottom row shows the trajectory of the tracked persons.

4.3 Online Robotic Evaluations

The online robotic evaluation corresponds to the experiments carried out on Rackham. As mentioned §4.1, the multi-modal detector and tracker are implemented in C/C++ embedded in GenoM modules. Both the detector and tracker run on the same computer while the LRF scan data is acquired through the second computer. The multi-modal detector alone runs from $1.5fps$ minimum to $4.5fps$ maximum depending on the number of hypothesis generated for the visual detector. The rate at which the combined system runs varies depending on the number of tracked persons and number of hypothesis generated by the laser for the visual detector. In our experiment, an approximate minimum of 0.7 frames per second was noted.

Recall the end goal is to realize a person following service robot with passer-by avoidance. The Person Following activity presented in (Germa et al., 2009) and depicted in the shaded area in figure 1 is based on an RFID system and a visual camera. A user (tagged person) wearing an RFID tag is tracked and followed by the robot irrespective of camera out of field of view, or occlusions. To check the integration of both systems, an experiment was carried out. In the experiment a tagged person is followed while a simple control law with rotative repulsive potential was used to avoid passers-by based on the multi-modal detector only. Figure 9 shows the a sequence of the video during a person following with obstacle avoidance based only on the multi-modal detector.

[2] A video of the tracking sequence is available at http://homepages.laas.fr/aamekonn/videos.htm

5 CONCLUSIONS

To conclude, this paper presented multi-modal person detection and tracking from a mobile robot based on LRF and vision intended for a socially acceptable navigation in crowded scenes during a person following activity. Though a person following scenario is considered, the framework is applicable for any service robot activity in a crowded public environment where perception of the whereabouts and dynamics of the persons around is required. It has been clearly shown that the multi-modal approach outperforms its single sensor counterparts taking detection, subsequent use, computation time, and precision all into account. Results obtained from offline and online robotic experiments have also been clearly reported asserting this statement.

Currently, investigations are on the way to use a LadyBug2 spherical camera to improve the detection and tracking further taking advantage of its wide field of view. Preliminary investigations are also underway with navigational schemes that consider the spatio-temporal information provided by our multi-target tracker.

REFERENCES

Alami, R., Chatila, R., Fleury, S., Ghallab, M., and Ingrand, F. (1998). An architecture for autonomy. *In Int.Journal of Robotics Research (IJRR'98)*, 17:315–337.

Breitenstein, M., Reichlin, F., Leibe, B., Koller-Meier, E., and Van Gool, L. (2009). Robust tracking-by-detection using a detector confidence particle filter. In *IEEE 12th Int. Conf. on Computer Vision (ICCV'09)*, pages 1515 –1522.

Calisi, D., Iocchi, L., and Leone, G. R. (2007). Person following through appearance models and stereo vision using a mobile robot. In *Proceedings of Int. Workshop on Robot Vision*, pages 46–56.

Chen, Z. and Birchfield, S. T. (2007). Person following with a mobile robot using binocular feature-based tracking. In *Proceedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS'07)*.

Cui, J., Zha, H., Zhao, H., and Shibasaki, R. (2005). Tracking multiple people using laser and vision. In *Proceedings of the 2005 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS'05)*, pages pp.1301–1306.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the Int. Conf. on Computer Vision and Pattern Recognition (CVPR'05)*, pages 886–893.

Felzenszwalb, P. F., Girshick, R. B., and McAllester, D. (2009). Discriminatively trained deformable part models, release 3. http://people.cs.uchicago.edu/ pff/latent-release3/.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D. A., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI'10)*, 32(9):1627–1645.

Fong, T. W., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*.

Germa, T., Lerasle, F., Ouadah, N., Cadenat, V., and Devy, M. (2009). Vision and RFID-based person tracking in crowds from a mobile robot. In *Proceedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS'09)*, pages 5591–5596. IEEE Press.

Isard, M. and MacCormick, J. (2001). Bramble: a bayesian multiple-blob tracker. In *Proceedings of 8th IEEE Int. Conf. on Computer Vision (ICCV'01)*, volume 2, pages 34 –41 vol.2.

Khan, Z., Balch, T., and Dellaert, F. (2005). Mcmc-based particle filtering for tracking a variable number of interacting targets. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI'05)*, 27(11):1805–1918.

Laptev, I. (2006). Improvements of object detection using boosted histograms. In *Proceedings of the British Machine Vision Conference (BMVC'06)*, pages 949–958.

Rasmussen, C. and Hager, G. D. (2001). Probabilistic data association methods for tracking complex visual objects. *IEEE Trans. Pattern Anal. Mach. Intell.(TPAMI'01)*, 23(6):560–576.

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

Schiele, B., Andriluka, M., Majer, N., Roth, S., and Wojek, C. (2009). Visual people detection: Different models, comparison and discussion. In *Proceedings of the IEEE ICRA 2009 Workshop on People Detection and Tracking*, pages 1–8.

Smith, K., Gatica-Perez, D., and Odobez, J.-M. (2005). Using particles to track varying numbers of interacting people. In *Proceedings of the Int. Conf. on Computer Vision and Pattern Recognition (CVPR'05)*, pages 962–969, Washington, DC, USA. IEEE Computer Society.

Spinello, L., Triebel, R., and Siegwart, R. (2008). Multimodal people detection and tracking in crowded scenes. In *Proceedings of the 23rd National Conference on Artificial intelligence (AAAI'08)*, pages 1409–1414. AAAI Press.

Xavier, J., Pacheco, M., Castro, D., and Ruano, A. (2005). Fast line, arc/circle and leg detection from laser scan data in a player driver. In *Proceedings of the Int. Conf. on Robotics and Automation (ICRA'05)*.

Zivkovic, Z. and Kröse, B. (2007). Part based people detection using 2d range data and images. In *Proceedings of the 2007 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS'07)*.