Vision and RFID Data Fusion for Tracking People in Crowds by a Mobile Robot

T.Germa[a,b], F.Lerasle[a,b], N.Ouadah[a,c], V.Cadenat[a,b]

{tgerma, lerasle, nouadah, cadenat}@laas.fr

[a] *CNRS ; LAAS ; 7, avenue du Colonel Roche, F-31077 Toulouse*
[b] *Université de Toulouse ; UPS, INSA, INP, ISAE ; LAAS-CNRS : F-31077 Toulouse*
[c] *CDTA/ENP; Cité 20 août 1956, Baba Hassen, Alger*

Abstract

In this paper, we address the problem of realizing a human following task in a crowded environment. We consider an active perception system, consisting of a camera mounted on a pan-tilt unit and a 360° RFID detection system, both embedded on a mobile robot. To perform such a task, it is necessary to efficiently track humans in crowds. In a first step, we have dealt with this problem using the particle filtering framework because it enables the fusion of heterogeneous data, which improves the tracking robustness. In a second step, we have considered the problem of controlling the robot motion to make the robot follow the person of interest. To this aim, we have designed a multi-sensor-based control strategy based on the tracker outputs and on the RFID data. Finally, we have implemented the tracker and the control strategy on our robot. The obtained experimental results highlight the relevance of the developed perceptual functions. Possible extensions of this work are discussed at the end of the article.

Key words: Radio frequency ID, multimodal data fusion, particle filtering, person tracking, person following.

1. Introduction

Giving a mobile robot the ability of automatically following a person appears to be a key issue to make it efficiently interact with humans. Numerous applications would benefit from such a capability. Service robotics is obviously one of these applications, as it requires interactive robots [**?**] able to follow a person to provide continual assistance in office buildings, museums,

Preprint submitted to CVIU *28th February 2013*

hospital environments, or even in shopping centers. Service robots clearly need to move in ways that are socially suitable for people. Such a robot have to localize its user, to discriminate him/her from others passers-by and to be able to follow him/her accross complex human-centered environment. In this context, tracking a given person in crowds from a mobile platform appears to be fundamental. However, numerous difficulties arise: moving cameras with limited view-field, cluttered background, illumination variations, hard real-time constraints, and so on.

The literature offers many tools to go beyond these difficulties. Our paper focuses on particle filtering framework as it easily enables to fuse heterogeneous data from embedded sensors. Despite their sporadicity, these dedicated person detectors and their hardware counterpart are very discriminant when present.

The paper is organized as follows. Section 2 depicts an overview of the corresponding works done within our robotic context and introduces our contributions. Section 3 describes our omnidirectional RFID prototype. This sensor is very discriminant when present in order to detect the user wearing an RFID tag. Section **??** recalls some PF basics and details our new importance function for multimodal person tracking. The developed control strategy to achieve a person following task in a crowded environment is detailed in section **??**, while section **??** presents the mobile robot which has been used for our tests and the obtained results. Finally, section **??** summarizes our contributions and discusses future extensions.

2. Overview and related work

Particle filters (PF) [**?**] through different schemes are currently investigated for person tracking in both robotics and vision communities. Besides the well-known CONDENSATION scheme, the fairly seldom exploited ICONDENSATION [**?**] variant steers sampling towards state space regions of high likelihood by incorporating both the dynamics and the measurements in the importance function. PF represent the posterior distribution by a set of samples, or particles, with associated importance weights. This weighted particles set is first drawn from an importance function and the state vector initial probability distribution, and is then updated over time taking into account the measurement models. Some approaches *e.g.* [**?**] show that intermittent and discriminant cues based on person detection and recognition functionalities must be considered in the importance function in order to: (i) automatically re-initialize the tracker on the targeted person when failures occur, (ii) simplify the data association problem in populated settings [?].

Primary, embedded detectors are generally restricted to stationary robots in order to (only) segment moving people from the background [? ?]. Some works [? ? ?] consider foreground segmentation based on disparity maps given a stereoscopic head [?], but they generally require significant CPU resources. Other techniques assume that people coarsely face the robot. In those cases, face detection [? ? ?] can be applied to (re)-initialize successfully the tracker after temporary occlusions, out of camera view-field, or target losses. These multi-view face detectors have received an increasing interest due to their computational efficiency. Such detectors have been extended to the full or upper human body detection [? ? ?]. Some complementary approaches combine person detection and recognition [? ?] in order to distinguish the targeted person from others. Nevertheless, despite many advances, a major problem - sensitivity to pose and illumination - still exists and a complete and on-boarded reliable visual-based solution that can be used in general conditions is not currently available. Clearly, using on-boarded monocular system to sense humans is very challenging compared to static and deported systems. Thus, face detection and skin color detection are only available when the person faces towards the robot, and the robot can hardly follow behind or even walk next to the person. Full or upper human body detectors based on supervised learning are inappropriate to cover all the person's range (from to 0.5 to 5m) and orientation[1] encountered when sensing from a mobile robot. Consequently, recent trends lead to methods based on multimodal sensor fusion. Their issue is generally to use the video stream as the primary sensor and other sensor streams as secondary ones.

Beyond visible spectrum vision, **thermal vision** allows to overcome some of the aforementioned limitations, since humans have a distinctive thermal profile with respect to non-living objects. Moreover, their appearance does not depend anymore on light conditions Yet, up to now, there are very few published works on using both thermal and visible cameras on mobile robots to detect/track humans (see a survey on thermal vision [?]). We can here mention the well-known PF proposed by Cielniak *et al.* in [?] which uses thermal vision for human detection and color images for capturing the appearance. Unfortunately, in crowds, sensing with thermal cameras leads

[1] The person can walk towards, away from, or past the robot, side-by-side, etc.

to an abundance of additional hot spots. It is then impossible to identify a given person as all humans (and also all living objects...) stick out as white regions on a black background.

Some other multimodal systems devoted to person tracking utilize **audio and visual sensors** [? ? ? ?]. In crowds, the data association problem can be settled by speaker identification [? ?]. Nevertheless, sensing people with audio cues during the robot or the customers movement is questionable. Indeed, the variability generated by the speaker, the recording conditions, the background noise especially in crowded environments, the inherent intermittence of the voice stream (as humans do not babble all the time) are the main difficulties which have to be overcome. Therefore, the speaker identification problem appears to be a challenging problem and still remains an open issue for further research.

Using **laser range finders** for person tracking is also frequent in the robotics community. In contrast to cameras, lasers provide accurate depth information, require little processing and are insensitive to ambient lighting. The classical strategy consists in extracting legs from a 2D laser scan at a fixed height. To this end, two particular types of features are intensively studied: motion [? ?] and geometry [? ? ? ? ?] features. Many multisensor fusion systems integrate the data provided by a laser range finder and a perspective [? ? ?] or omnidirectional [? ?] camera. Anyway, systems involving laser scans suffer from several drawbacks. Leg detection in a 2D scan does not provide robust features for discriminating the different persons in the robot vicinity, while the detector fails when one leg occludes the other.

Recent person tracking approaches have focused on **indoor positioning systems** based on wireless networking indoor infrastructure and ultrasound, infrared [?], or radio frequency badged humans' clothes [? ? ? ? ?]. Radio frequency (RF) signals are widely used as they (i) can penetrate through most of the building material, (ii) have an excellent range in indoor environments, (iii) have less interferences with other frequency components. Moreover, RFID tags are preferred to accelerometers for aesthetical and ergonomical reasons [? ? ?]. Common applications involving RFID technologies [? ? ? ? ?] assume stationary readers distributed throughout the settings, namely ubiquitous sensors. Solely Schutz *et al.* in [?] considered the multimodal people tracking from a network of RF sensors and laser range finders placed throughout an environment. Our approach privileges on-board perceptual resources (monocular color vision and RF reader) in order to limit the hardware installation cost and therefore the indoor setting support. We can here mention the approach proposed in [?] which considers an on-board RF device for people detection. However, the detection range was limited to 180° and no multimodal data fusion was done.

RFID sensors enjoy the nice properties to provide explicit information about the person identity, even if the location information is relatively coarse. Our multimodal person tracker combines the accuracy and information richness advantages of active color vision with the identification certainty of RFID. This tracker, which has not been addressed in the literature, is expected to be more resilient to occlusions than vision-based only systems, since the former benefits from a coarse estimate of people location in addition to the knowledge of his/her appearance. Furthermore, the ID-sensor can act as reliable stimuli that triggers the vision system. Finally, when several people lie in the camera view-field[2], this multimodal sensor data fusion will help in distinguishing the targeted person from the others.

The contributions of the paper is three-fold. The first contribution of this paper is the customization of an off-the-shelf RFID system to make it able to detect tags in 360° view field, by multiplexing 8 antennas. We have embedded this system on our mobile robot Rackham to detect passive RFID-tagged persons. This omnidirectional ID-sensor, unaffected by lighting conditions or humans' appearance, appears as an ideal complement to trigger a PTU-mounted perspective camera. The second contribution concerns particle sampling within the ICONDENSATION scheme. We propose a genuine importance function based on probabilistic saliency maps issued from visual and RF person detector and ID as well as a rejection sampling mechanism to (re)-positions samples on the desired person during tracking. This particle sampling strategy, which is unique in the literature, should improve our multi-sensor based tracker so that it becomes much more resilient to occlusions, data association, and target losses than vision-based only systems. The last contribution concerns a multi-sensor based control making the mobile robot reliably follow in real time a person in a more difficult setting than other previous works [? ? ?].

[2] In this case, there are multiple observations in the image plane.

3. Person detection and identification based on RFID

3.1. Device description

The device consists of: (i) a CAENRFID[3] $A941$ multiprotocol off-the-shelf reader which works at 870MHz, with a programmable emitting RF power from 100 to 1200mW, (ii) 8 directive antennas to detect the passive tags worn on the customer's clothes, (iii) a prototype circuit in order to sequentially initialize each antenna (figure 1). With a single antenna, only a tag angle relative to the antenna plane can be estimated. With our 8 antennas, the tag can be detected all around the robot at any distance between 0.5m (*i.e.* approximately the robot's radius) and 5m. Given the placement of the antennas and their own field of view, the robot neighborhood is divided into 24 areas (figure **??**), depending on the number of antennas simultaneously detecting the RFID tag.

To determine the observation model of the whole antenna set, statistics are performed by counting frequencies depending on the number of antennas (three at a maximum, figure **??**) that detect the tag. The resulting normalized histograms are shown in figure **??** where the x-axis represents the azimuthal angle θ_{tag}. Similar histograms can be observed for the distance d_{tag}[4]. The resulting sensor model makes the simplifying assumption that both azimuth and distance histograms can be approximated by Gaussians respectively defined by $(\mu_{\theta_{tag}}, \sigma_{\theta_{tag}})$ and $(\mu_{d_{tag}}, \sigma_{d_{tag}})$ where $\mu_{(.)}$ and $\sigma_{(.)}$ are the mean and standard deviation. Afterwards, we project these probabilities for the current tag position to a saliency map of the floor. The size of the saliency map is 300×300 pixels; thus the area of each pixel represents 7 cm^2. Each pixel probability is calculated given the 8-antenna set outputs to approximate the RFID tag position (figure **??**). The three rightmost plots in figure **??** respectively shows the saliency maps

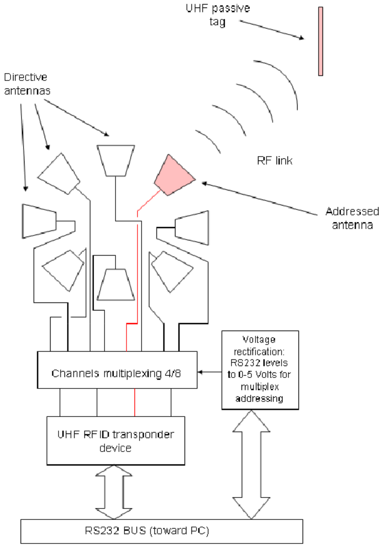
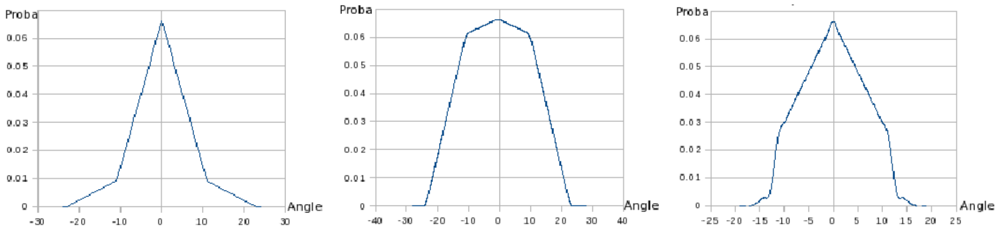

UHF passive tag
Directive antennas
RF link
Addressed antenna
Voltage rectification: RS232 levels to 0-5 Volts for multiplex addressing
Channels multiplexing 4/8
UHF RFID transponder device
RS232 BUS (toward PC)

Figure 1: RF multiplexing prototype to address 8 antennas.

[3] see http://www.caen.it/rfid/

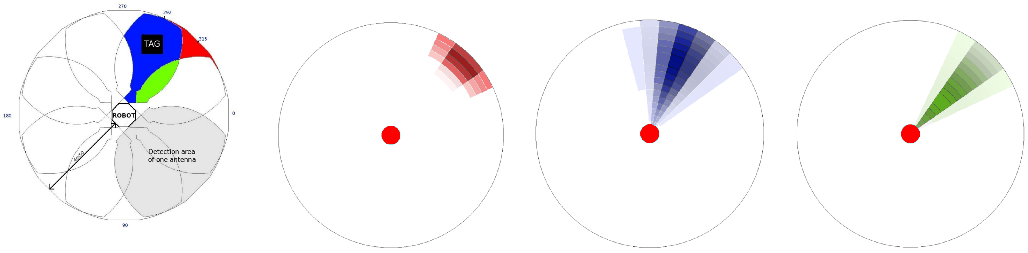
[4] They are not presented here to save space, but they are available on request.

for the detection by one, two or three antennas. Given this observation model, evaluations allow to characterize the ID-sensor performances.



(a) (b) (c)

Figure 2: Occurrence frequencies of angle θ_{tag} given one (a), two (b) or three (c) detections.



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

Figure 3: Azimuthal view field of 8 antennas (a) and saliency map for tag detection respectively for 1 (b), 2 (c) and 3 (d) antennas.

3.2. Evaluations from feasibility study

The RF system has been mounted on our mobile robot Rackham (section **??**) and evaluated in the presence of people. We have proceeded in the following way. We have generated statistics by counting frequencies on a 81m^2 area around the robot. Obstacles have been added one by one during the test runs. Their positions have been randomly chosen and uniformly distributed in this area. The corresponding ground-truth is based on the ratio between the occluding zones induced by obstacles (assuming an average person-width of 40cm) and the total area.

Given such various “crowdedness” situations, the RFID tag has been moved around the robot assuming no self-occlusion by the person wearing the tag during this evaluation. We have repeated this sequence for different distances and we have counted for every point in a discrete grid whether the tag worn by a fixed person is detected or not, depending on the crowdedness. Comparisons between experimental and theoretical detection rates are shown in figure **??** (see the box-and-whisker plots).

Figure 4: Detection rate versus crowdedness in the robot surrounding.

The x-axis and y-axis respectively denote the number of occluding persons (that is "crowdedness") and the detection rate. The box plots and the thick stretches inside indicate the degree of dispersion (for 50% of the trials) and the median of the trials. Our experimental curves are shown to be rather close to the theoretical ones. As the system is disturbed by the occlusions, the number of false-negative readings logically increases with the number of obstacles. Nevertheless, the detection rate remains satisfactory, even for overcrowded scenes (*e.g.*70% in average for 7 persons standing around the robot). Furthermore, very few false-positive readings (reflections, detections with the wrong antennas...) are observed in practice[5].

4. Person tracking using vision and RFID

4.1. Basics on particle filters and data fusion

Particle filters (PF) aim at recursively approximating the posterior probability density function (pdf) $p(\mathbf{x}_k|z_{1:k})$ of the state vector $\mathbf{x}_k$ at time k conditioned on the set of measurements $z_{1:k} = z_1, \ldots, z_k$. A linear point-mass combination

$$p(\mathbf{x}_k|z_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)} \delta(\mathbf{x}_k - \mathbf{x}_k^{(i)}),\ \sum_{i=1}^{N} w_k^{(i)} = 1, \tag{1}$$

is determined where $\delta(.)$ is the Dirac distribution. It expresses the selection of a value –or "particle"– $\mathbf{x}_k^{(i)}$ with probability –or "weight"– $w_k^{(i)}$, $i = 1, \ldots, N$. An approximation of the conditional expectation of any function of $\mathbf{x}_k$, such as the MMSE[6] estimate $\mathrm{E}_{p(\mathbf{x}_k|z_{1:k})}[\mathbf{x}_k]$, then follows.

[5] Passive tags induce few signal reflections contrary to their active counterparts.

[6] for ”Medium Mean Square Estimate“

Algorithm 1 Generic particle filtering algorithm (SIR).

```
Require: $[\{\mathbf{x}_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^{N}, z_k$
 1: if $k = 0$ then
 2:   Draw $\mathbf{x}_0^{(1)}, \dots, \mathbf{x}_0^{(i)}, \dots, \mathbf{x}_0^{(N)}$ i.i.d. according to $p(\mathbf{x}_0)$, and set $w_0^{(i)} = \frac{1}{N}$
 3: end if
 4: if $k \geq 1$ then {—$[\{\mathbf{x}_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^{N}$ being a particle description of $p(\mathbf{x}_{k-1}|z_{1:k-1})$—}
 5:   for $i = 1, \dots, 10$ do
 6:     "Propagate" the particle $\mathbf{x}_{k-1}^{(i)}$ by independently sampling $\mathbf{x}_k^{(i)} \sim q(\mathbf{x}_k|\mathbf{x}_{k-1}^{(i)}, z_k)$
 7:     Update the weight $w_k^{(i)}$ associated to $\mathbf{x}_k^{(i)}$ according to $w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(z_k|\mathbf{x}_k^{(i)})p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})}{q(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, z_k)}$,
 8:     Prior to a normalization step so that $\sum_i w_k^{(i)} = 1$
 9:   end for
10:   Compute the conditional mean of any function of $\mathbf{x}_k$, e.g. the MMSE estimate $\mathrm{E}_{p(\mathbf{x}_k|z_{1:k})}[\mathbf{x}_k]$,
      from the approximation $\sum_{i=1}^{N} w_k^{(i)} \delta(x_k - x_k^{(i)})$ of the posterior $p(x_k|z_{1:k})$
11:   At any time or depending on an "efficiency" criterion, resample the description $[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^{N}$
      of $p(\mathbf{x}_k|z_{1:k})$ into the equivalent evenly weighted particles set $[\{x_k^{(s^{(i)})}, \frac{1}{N}\}]_{i=1}^{N}$, by sampling in
      $\{1, \dots, N\}$ the indexes $s^{(1)}, \dots, s^{(N)}$ according to $P(s^{(i)} = j) = w_k^{(j)}$; set $\mathbf{x}_k^{(i)}$ and $w_k^{(i)}$ to $\mathbf{x}_k^{(s^{(i)})}$
      and $\frac{1}{N}$
12: end if
```

The "Sampling Importance Resampling" (SIR), shown on Algorithm **??**, is fully described by the prior $p(\mathbf{x}_0)$, the dynamics pdf $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ and the observation pdf $p(z_k|\mathbf{x}_k)$. After initialization of independent identically distributed (i.i.d.) sequence drawn from $p(\mathbf{x}_0)$, the particles stochastically evolve, being sampled from an importance function $q(\mathbf{x}_k|\mathbf{x}_{k-1}^{(i)}, z_k)$. They are then suitably weighted to guarantee the consistency of the approximation (**??**). To this end, step 7 assigns each particle $\mathbf{x}_k^{(i)}$ a weight $w_k^{(i)}$ involving its *likelihood* $p(z_k|\mathbf{x}_k^{(i)})$ w.r.t. the measurement z_k as well as the values of the dynamics pdf and importance function at $\mathbf{x}_k^{(i)}$. In order to limit the well known degeneracy phenomenon [**?**], step 11 inserts a resampling stage introduced by Gordon *et al.* in [**?**] so that the particles associated with high weights are duplicated while the others collapse and the resulting sequence $\mathbf{x}_k^{(s^{(1)})}, \dots, \mathbf{x}_k^{(s^{(N)})}$ is i.i.d. according to (**??**).

The CONDENSATION –for "Conditional Density Propagation" [**?**]– is the instance of the SIR algorithm such that the particles are drawn according to the system dynamics, viz. when $q(\mathbf{x}_k|\mathbf{x}_{k-1}^{(i)}, z_k) = p(\mathbf{x}_k|\mathbf{x}_{k-1}^{(i)})$. Then, in visual tracking, the original algorithm [**?**] defines the particles likelihoods from contour primitives, but other visual cues have also been exploited [**?**]. On this point, resampling may lead to a loss of diversity in the state space exploration. The importance function must thus be carefully defined. As CONDENSATION draws the particles $\mathbf{x}_k^{(i)}$ from the system dynamics but "blindly" w.r.t. the measurement z_k, many of them may be assigned a low likelihood $p(z_k|\mathbf{x}_k^{(i)})$ and thus a low weight in step 7, which significantly worsen the overall filter performance.

An alternative, henceforth labeled "Measurement-based SIR" (MSIR), merely consists in sampling the particles – or just some of their entries – at time k according to an importance function $\pi(\mathbf{x}_k|z_k)$ defined from the current image. The first MSIR strategy was ICONDENSATION [**?**], which guided the state space exploration by a color blob detector. Other visual detection functionalities can be used as well, *e.g.* face detection/recognition (see below), or any other intermittent primitive which, despite its sporadicity, is very discriminant when present [**?**]. Thus, the classical importance function $\pi(.)$ based on a single detector can be extended to consider the outputs from L detection modules, *i.e.*

$$\pi(\mathbf{x}_k^{(i)}|z_k^1,\ldots,z_k^L) = \sum_{l=1}^{L} \kappa_l \pi(\mathbf{x}_k^{(i)}|z_k^l),\ \textit{with} \sum \kappa_l = 1. \qquad (2)$$

In an MSIR scheme, if a particle $\mathbf{x}_k^{(i)}$ drawn exclusively from the image (namely $\pi(.)$) is inconsistent with its predecessor $\mathbf{x}_{k-1}^{(i)}$ from the point of view of the state dynamics, the update formula leads to a small weight $w_k^{(i)}$. One solution to this problem, as proposed in the genuine ICONDENSATION algorithm, consists in also sampling some particles from the dynamics and some w.r.t. the prior so that:

$$q(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, z_k) = \alpha\pi(\mathbf{x}_k^{(i)}|z_k) + \beta p(\mathbf{x}_k|\mathbf{x}_{k-1}^{(i)}) + (1-\alpha-\beta)p_0(\mathbf{x}_k). \qquad (3)$$

with $\alpha, \beta \in [0;1]$. Besides the importance function, the measurement function involves visual cues which must be persistent but are more prone to ambiguity for cluttered scenes. An alternative is to consider multi-cue fusion in the weighting stage. Given L measurement sources $(z_k^1,\ldots,z_k^L)$ and assuming the latter are mutually independent conditioned on the state, the unified measurement function can then be factorized as follows:

$$p(z_k^1,\ldots,z_k^L|\mathbf{x}_k^{(i)}) \propto \prod_{l=1}^{L} p(z_k^l|\mathbf{x}_k^{(i)}). \qquad (4)$$

4.2. Tracking implementation

The aim is to fit the template relative to the RFID-tagged person all along the video stream through the estimation of his/her image coordinates (u, v) and its scale factor s of his/her head. All these parameters are accounted for in the above state vector $\mathbf{x}_k$ related to the k-th frame. With regard to the dynamics $p(\mathbf{x}_k|\mathbf{x}_{k-1})$, the image motions of humans are difficult to characterize over time. This weak knowledge is modelled by defining the state vector as $\mathbf{x}_k = [u_k, v_k, s_k]'$ and assuming that its entries evolve according to mutually independent random walk models, viz. $p(\mathbf{x}_k|\mathbf{x}_{k-1}) = \mathcal{N}(\mathbf{x}_k; \mathbf{x}_{k-1}, \Sigma)$ where $\mathcal{N}(.; \mu, \Sigma)$ is a Gaussian distribution with mean μ and covariance $\Sigma = \text{diag}(\sigma_u^2, \sigma_v^2, \sigma_s^2)$.

In both importance sampling and weight update steps, fusing multiple cues enables the tracker to better benefit from distinct information, and decrease its sensitivity to temporary failures in some of the measurement processes. The underlying unified likelihood in the weighting stage is more or less conventional. It is computed thanks to (**??**) by means of several measurement functions, according to persistent visual cues, namely: (i) edges to model the silhouette [**?**], (ii) multiple color distributions to represent the person's appearance (both head and torso) [**?**]. Despite its simplicity, our measurement function is inexpensive while still providing some level of person discrimination from the clothes appearance. Otherwise, our importance function is unique in the literature and is detailed below.

4.3. Importance function based on visual and RF cues

Given equation (**??**), three functions $\pi(\mathbf{x}_k|z_k^c)$, $\pi(\mathbf{x}_k|z_k^s)$ and $\pi(\mathbf{x}_k|z_k^r)$, respectively based on skin probability image, face detector and RF identification are considered.

The importance function $\pi(\mathbf{x}_k|z_k^c)$ at location $\mathbf{x}_k = (u, v)$ is described by $\pi(\mathbf{x}|z^c) = \mathbf{h}(c_z(\mathbf{x}))$ where $c_z(\mathbf{x})$ is the color of the pixel located in $\mathbf{x}$ in the input image z^c. $\mathbf{h}$ is the 3D normalized histogram used for backprojection [**?**] indexed by R, G, B channels and represents the color distribution of the skin which is *a priori* learnt.

The function $\pi(\mathbf{x}_k|z_k^s)$ relies on a probabilistic image based on the well-known face detector pioneered by Viola *et al.* in [**?**], and improved in [**?** **?**], which covers a range of ±45° out-of plane rotation. Let N_B be the number of detected faces $\{\mathcal{F}_j\}_{j=1}^{N_B}$ and $\mathbf{p}_i = (u_i, v_i), i = 1, \ldots, N_B$ the centroid coordinate of each such region. The face recognition technique, detailed in [**?**], involves two steps during the learning stage. The first one is composed of PCA-based computation and multi-class SVM[7] learning, while the second one uses a genetic algorithm free-parameters optimization based on NSGA-II. Finally, our on-line decision rule defines a posterior probability $P(C_t|\mathcal{F}, z)$ of labeling face $\mathcal{F}_j$ to C_t so that:

$$\begin{cases} \forall t \; P(C_t|\mathcal{F}, z) = 0 \text{ and } P(C_\emptyset|\mathcal{F}, z) = 1 \text{ when } \forall t \; \mathscr{L}^t < \tau \\ \forall t \; P(C_t|\mathcal{F}, z) = \frac{\mathscr{L}^t}{\sum_p \mathscr{L}^p} \text{ and } P(C_\emptyset|\mathcal{F}, z) = 0 \text{ otherwise} \end{cases},$$

where $C_\emptyset$ refers to the void class, τ is one of the free-parameters of the system and C_t refers to the face basis of the RFID-tagged person. The function $\pi(.)$ at location $\mathbf{x} = (u, v)$ is deduced using a weighted Gaussian mixture proposal[8]. Its expression is given hereafter:

$$\pi(\mathbf{x}|z^s) \propto \sum_{j=1}^{N_B} P(C|\mathcal{F}_j, z).\mathcal{N}(\mathbf{x}; \mathbf{p}_j, \mathrm{diag}(\sigma^2_{u_j}, \sigma^2_{v_j})),$$

where $P(C|\mathcal{F}_j, z)$ is the face ID probabilities for each detected face $\mathcal{F}_j$ given beforehand learnt tracked person face. Given the RF outputs, the function $\pi(\mathbf{x}_k^{(i)}|z_k^r)$ expresses as follows:

$$\pi(\mathbf{x}_k^{(i)}|z^r) = \mathcal{N}(\theta_{\mathbf{x}_k^{(i)}}; \mu_{\theta_{tag}}, \sigma_{\theta_{tag}}),$$

where $\theta_{\mathbf{x}_k^{(i)}}$ is the azimuthal position of the particle $\mathbf{x}_k^{(i)}$ in the robot frame, deduced from its horizontal position on the image and the camera pan angle. $\mu_{\theta_{tag}}$ and $\sigma_{\theta_{tag}}$, described in section 3, are respectively the mean and the covariance of the estimated position of the sole targeted tag associated to the user in the robot frame depending on the antenna outputs.

The particle sampling is done using the importance function $q(.)$ given in equation (**??**) and requires a process of rejection sampling. This process constitutes an alternative when $q(.)$ is not analytically modelled. The principle is described in algorithm **??** with $g(.)$ an instrumental distribution to make the sampling easier under the restriction that $q(.) < Mg(.)$ where $M > 1$ is an appropriate bound on $\frac{q(.)}{g(.)}$.

Figure **??** shows an illustration of the rejection sampling algorithm for a given image. Our importance function (**??**) combined with rejection sampling ensures that the particles will be placed in the relevant areas of the state space *i.e.* concentrated on the tracked person or potential candidate areas.

[7] for "Support Vector Machine"

[8] Indexes k and i are omitted for the sake of clarity and space.

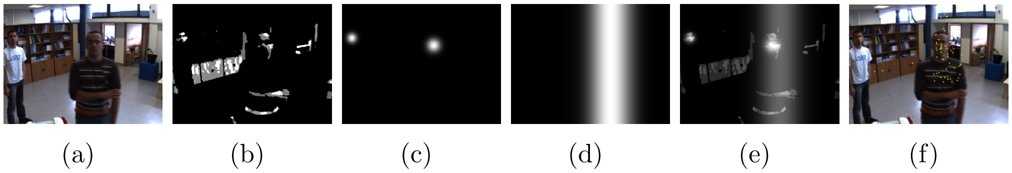


Figure 5: (a) original image, (b) skin probability image $\pi(\mathbf{x}_k|z_k^c)$, (c) face detection $\pi(\mathbf{x}_k|z_k^s)$, (d) azimuthal angle from RFID detection $\pi(\mathbf{x}_k|z_k^r)$, (e) unified importance function (**??**) (without dynamic), (f) accepted particles (yellow dots) after rejection sampling.

Algorithm 2 Rejection sampling algorithm.

```
draw $\mathbf{x}_k^{(i)}$ according to $Mg(\mathbf{x}_k)$
$r \leftarrow \frac{q(\mathbf{x}_k|\mathbf{x}_{k-1},z_k)}{Mg(\mathbf{x}_k^{(i)})}$
draw $u$ according to $\mathcal{U}_{[0,1]}$
if $u \leq r$ then
    accept $\mathbf{x}_k^{(i)}$
else
    reject it
end if
```

5. A sensor-based control law for person following task

Now, we address the problem of making the robot follow the tagged person. To this aim, we use the data provided by both the tracker and the RFID system. We first briefly present the considered robotic system and the chosen control strategy, before detailing the different designed control laws.

5.1. Modelling the problem: the robot and the control strategy

Our robot Rackham depicted in section **??** consists of a nonholonomic mobile base equipped with a RFID system and with a camera mounted on a pan/tilt unit (PTU). Four control inputs can then be used to act on our robot: the linear and angular mobile base velocities (v_r, ω_r) and the pan/tilt unit angular velocities (ω_p, ω_t). Our goal is to compute these four velocities so that the robot can efficiently and safely achieve the person following. Different control strategies are available in the literature. In our case where the camera and the RFID tag are used to detect and track the user, it appears rather natural to consider visual servoing techniques [**? ?**]. These techniques allow to control a large panel of robotic systems using image data provided