The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Vision and RFID-based Person Tracking in Crowds from a Mobile Robot

Thierry Germa[1,2], Frédéric Lerasle[1,2], Noureddine Ouadah[1,3], Viviane Cadenat[1,2], Michel Devy[1,2]

***Abstract*— This paper deals with the tracking of persons in a human cluttered environment. It is performed by an active perception system, consisting of a camera mounted on a pan-tilt unit and a 360° RFID detection system which are embedded on a mobile robot. Particle filters enable the fusion of heterogeneous data into the proposal distribution from which the particles are sampled. The information provided by the tracker is then used to build sensor-based dedicated control laws in order to make the robot follow the RFID tagged person. Finally, experiments on our mobile robot are presented in order to highlight the relevance and complementarity of the developed perceptual functions.**

I. Introduction and framework

The ability for a mobile robot to automatically follow a person in public areas is a key issue to effectively interact with the surrounding world. To fulfill this objective, robust algorithms able to track a given person thanks to multiple on-boarded sensors are required. Particle filters [2] are currently investigated for person tracking in both robotics and vision communities. Their popularity stems from their ability to fuse in a probabilistic way various kinds of visual measurements. Pérez *et al.* in [10] highlight the fact that intermittent cues are suitable candidates for the construction of detection modules and efficient proposal distributions. Clearly, reliable people detectors improve the tracking performance.

Visual detectors commonly assume that people face toward the robot, so that methods based on face detection/recognition [5], [12] can be applied to successfully (re)-initialize successfully the tracker after temporary occlusion, camera out view-field, or target loss. Their performances heavily depend on the light conditions, viewing angle, distance to persons, and variability of humans' appearance in video streams. Consequently, numerous approaches consider multimodal sensor fusion. Their issue is to combine different sensor streams from microphones [10], laser range finders [13] with the video stream.

Recent approaches have also focused on "emerging technologies" based on both wireless networking indoor infrastructure and ultrasound, infrared [11], or radio frequency (RF) badges worn on clothes [8].

Such passive ID-sensors provide explicit information about the person ID, even if the location information is relatively coarse. Our system combines the accuracy and information richness advantages of active color vision with the identification certainty of RFID and has not been addressed in the literature yet. This tracker 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, this multimodal sensor data fusion will simplify the data association problem.

To our knowledge, only Hahnel *et al.* in [6] has considered an on-board RF device for people detection. However, the detection range was limited to 180° angle and a sole sensor is used: the RFID tag. Common applications involving RFID technologies assume stationary readers distributed throughout the settings [8], [11], namely ubiquitous sensors. Our approach privileges on-board perceptual resources in order to limit the hardware installation cost and therefore the indoor setting support. Using visual and RF detectors into the proposal distribution of the PF, should improve our tracker so that it becomes much more resilient to occlusions, data association, and target loss than vision-based only systems. Our multimodal tracker allows to reliably follow a person in a complex dynamic, crowded environments.

The paper is organized as follows. Section II depicts our omnidirectional RFID prototype. Section III recalls some PF basics and details our importance function for multimodal person tracking. Section IV details the developed control strategy to achieve a person following task in a crowded environment. Section V presents our robotic platform and associated live experiments. Finally, section VI summarizes our contributions and discusses future extensions.

II. Human user detection based on RFID

A. Device description

The device consists of: (i) a CAENRFID $A941$ off-the-shelf multiprotocol reader working at 870MHz, (ii) 8 directive antennas to detect the passive tags worn by the customer, (iii) a prototype circuit in order to sequentially adress each antenna (figure 1). With a single antenna, only a tag angle relative to the antenna plane can be estimated. Thus, with our 8 antennas, the tag can be detected all around the robot.

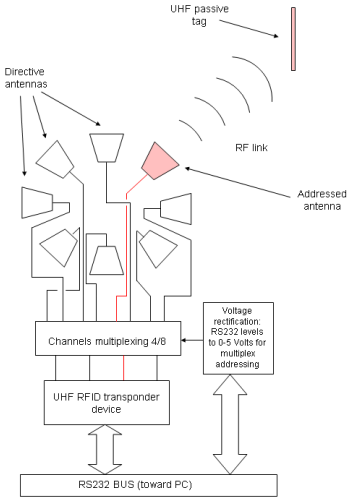


Fig. 1. RF multiplexing prototype to adress 8 antennas.

Given the placement of the antennas and their own field of view, the robot surrounding is divided into 24 areas (figure 3(a)), depending on the number of antennas detecting simultaneously the RFID tag. This lattest can be detected by the reader all around the robot at any distance between 0.5m (*i.e.* approximately the robot's radius) and 4.5m. To determine the observation model for the whole antenna set, statistics are performed by counting frequencies depending on the number (three at a maximum, figure 3(a)) of antennas that detect the same tag.

1 CNRS ; LAAS ; 7, avenue du Colonel Roche, F-31077 Toulouse, France
2 Université de Toulouse ; UPS, INSA, INP, ISAE ; LAAS-CNRS : F-31077 Toulouse, France
3 CDTA/ENP; Cité 20 août 1956, Baba Hassen, Alger, Algérie
firstname.lastname@laas.fr

The resulting normalized histograms are shown in figure 2 where the x-axis represents the azimuthal angle (denoted by θ)[1]. The resulting sensor model assumes that both azimuth and distance histograms can be approximated by Gaussians respectively defined by the couples $(\mu_\theta^{tag}, \sigma_\theta^{tag})$ and $(\mu_d^{tag}, \sigma_d^{tag})$ where $\mu_{(.)}^{tag}$ and $\sigma_{(.)}^{tag}$ are the mean and standard deviation. Afterwards, we project these probabilities for the current tag position to a saliency map of the floor. Each pixel probability is calculated given the antenna outputs to achieve approximate position estimation of the RFID tag (figure 3). Given this model observation, evaluations allow to characterize the ID-sensor performances.

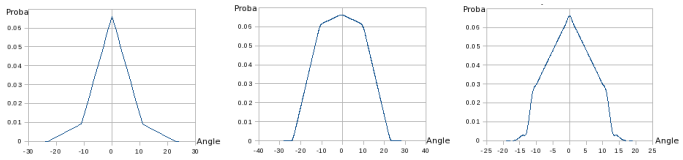
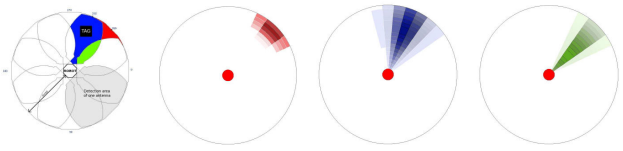


Fig. 2. From left to right: occurrence frequencies of azimuthal angle given one, two or three detections.



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

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

B. Evaluations from feasibility study

The RF system has been mounted on our robot Rackham, an iRobot B21r mobile platform and evaluated in a crowded context. We have generated a statistics by counting frequencies on a 81m^2 area around the robot. Obstacles with random positions uniformly distributed in this area are added one by one during the test runs. The corresponding ground-truth is based on the ratio between the areas occluding the tag induced by obstacles (assuming an average person-width of 40cm) and the total area. Given such various "crowdedness" situations, the RFID tag is moved around the robot. We repeat this experimental protocol for different distances and count 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 reported in figure 4 thanks to box-and-whisher plots. The x-axis and y-axis denote the "crowdedness" and the detection rate while the box plots and the thick stretches inside indicate the degree of dispersion (for 50% of the trials) and median of the trials. Our experimental curves are shown to be closed to the theoretical ones. As the system is disturbed by 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 around the robot). Furthermore, very few false-positive readings (reflections, detections with the wrong antennas...) are observed in practice[2].

[1]Histograms relatively to the tag/reader distances (noted d) are not presented for space reasons but they are available on request.

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

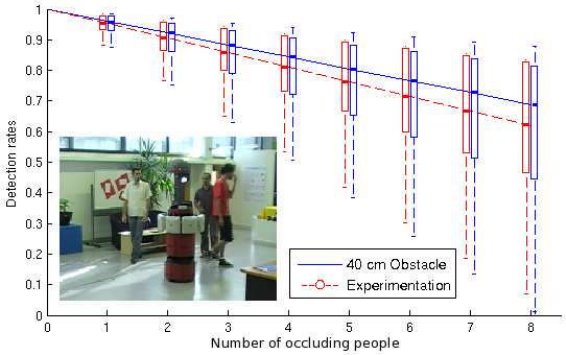


Fig. 4. Detection rate versus crowdedness in the robot surrounding.

III. Person tracking using vision and RFID

A. Basics about particle filters

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

$$p(x_k|z_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)} \delta(x_k - x_k^{(i)}), \ \sum_{i=1}^{N} w_k^{(i)} = 1, \qquad (1)$$

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

Recall that the SIR –or "Sampling Importance Resampling" – algorithm is fully described by the prior $p(x_0)$, the dynamics pdf $p(x_k|x_{k-1})$ and the observation pdf $p(z_k|x_k)$. After initialization of independent identically distributed (i.i.d.) sequence drawn from $p(x_0)$, the particles evolve stochastically, being sampled from an importance function $q(x_k|x_{k-1}^{(i)}, z_k)$. They are then suitably weighted so as to guarantee the consistency of the approximation (1). Then a weight $w_k^{(i)}$ is affected to each particle $x_k^{(i)}$ involving its likelihood $p(z_k|x_k^{(i)})$ w.r.t. the measurement z_k as well as the values of the dynamics pdf and importance function at $x_k^{(i)}$. In order to limit the well-known degeneracy phenomenon [2], a resampling stage is introduced so that the particles associated with high weights are duplicated while the others collapse and the resulting sequence $\{x_k^{(s^{(i)})}\}_{i=1}^N$ is i.i.d. according to (1).

With respect to our heterogenous data fusion context, we have chosen to use ICONDENSATION [7], that consists in sampling some particles from the observation (namely $\pi(.)$), some from the dynamics and some w.r.t. the prior so that the importance function is given by:

$$q(x_k^{(i)}|x_{k-1}^{(i)}, z_k) = \alpha\pi(x_k^{(i)}|z_k) + \beta p(x_k^{(i)}|x_{k-1}^{(i)}) + (1-\alpha-\beta)p_0(x_k^{(i)}). \qquad (2)$$

with $\alpha, \beta \in [0;1]$. $\pi(.)$ relates to detector outputs which, despite their intermittent nature, are proved to be very discriminant when present [10].

B. Tracking implementation

The aim is to fit a 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 motions of humans in an image are difficult to characterize over time. This weak knowledge is formalized by defining the state vector as $\mathbf{x}_k = [u_k, v_k, s_k]'$ and assuming that it evolves 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 $\Sigma = \mathrm{diag}(\sigma_x^2, \sigma_y^2, \sigma_s^2)$.

In importance sampling, fusing multiple cues enables the tracker to better benefit from distinct information, and to decrease its sensitivity to temporary failures in some of the measurement processes. The underlying unified likelihood $p(z_k|x_k^{(i)})$ in the weighting stage is more or less conventional. It is computed by means of several measurement functions, according to persistent visual cues, namely: (i) multiple color distributions to represent the person's appearance (both head and torso), (ii) edges to model the silhouette. The reader is referred to [5] for more details. Otherwise, our importance function is unique in the literature and so is detailed here below.

C. Importance function based on visual and RF cues

Recall that the function $\pi(.)$ in equation (2) offers a mathematically principled way of directing search according to multiple and possibly heterogeneous detectors. Given L independent detectors and κ their predefined weights, the function $\pi(.)$ can be reformulated as

$$\pi(x_k^{(i)}|z_k^1, \ldots, z_k^L) = \sum_{l=1}^{L} \kappa_l . \pi(x_k^{(i)}|z_k^l), \text{ with } \sum \kappa_l = 1. \quad (3)$$

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 [9], face detector and RF identification are here considered.

The importance function $\pi(\mathbf{x}_k|z_k^c)$ at location $\mathbf{x}_k = (u, v)$ is described by

$$\pi(\mathbf{x}|z^c) = \mathbf{h}(c_z(\mathbf{x})) \quad (4)$$

given that $c_z(\mathbf{x})$ is the color of the pixel situated in $\mathbf{x}$ in the input image z^c and $\mathbf{h}$ is the normalized histogram representing the color distribution of the skin *a priori* learnt. The function $\pi(\mathbf{x}_k|z_k^s)$ is based on a probabilistic image based on the well-known face detector pioneered by Viola *et al.* in [12]. Let N_B be the number of detected faces and $\mathbf{p}_i = (u_i, v_i), i = 1, \ldots, N_B$ the centroid coordinate of each such region. The function $\pi(.)$ at location $\mathbf{x} = (u, v)$ follows, as the Gaussian mixture proposal[3]

$$\pi(\mathbf{x}|z^s) \propto \sum_{j=1}^{N_B} P(C|\mathcal{F}_j, z) . \mathcal{N}(\mathbf{x}; \mathbf{p}_j, \mathrm{diag}(\sigma_{u_j}^2, \sigma_{v_j}^2)), \quad (5)$$

with $P(C|\mathcal{F}_j, z)$ the face ID probabilities for each detected face $\mathcal{F}_j$ given previously learnt eigenfaces of the tracked person. The technique, which is out of the paper scope, is detailed in [4]. Concerning the RF outputs, the function $\pi(\mathbf{x}_k|z_k^r)$ follows

$$\pi(\mathbf{x}|z^r) = \mathcal{N}(\theta_{\mathbf{x}}; \mu_\theta^{tag}, \sigma_\theta^{tag}), \quad (6)$$

given that $\theta_{\mathbf{x}}$ is the azimuthal position of the particle $\mathbf{x}$ in the robot frame, deduced from its horizontal position on the image and the camera pan angle, and μ_θ^{tag} and σ_θ^{tag} are respectively the mean and the covariance of the estimated position of the tag in the robot frame depending on the antenna outputs and described in section II.

[3]Index k and (i) are omitted for the sake of clarity and space.

The particle sampling is done using the importance function $q(.)$ in equation (2) and a process of rejection sampling. This process constitutes an alternative when $q(.)$ is not analytically described. The principle is described in algorithm 1 with $g(.)$ an instrumental distribution to make the sampling easier under the restriction that $q(.) < Mg(.)$ where $M > 1$ is an appropriate bound on $\frac{q(.)}{g(.)}$[4].

Algorithm 1 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

Figure 5 shows an illustration of the rejection sampling algorithm described for a given image. Our importance function (2) 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 canditate areas.

IV. A SENSOR-BASED CONTROL LAW FOR PERSON FOLLOWING TASK

Here, our idea is to benefit from the information provided by the previously developed multimodal tracker to build a control strategy able to achieve a particular mission consisting in following the user thanks to his image coordinate.

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

Our robot Rackham (depicted in section V) consists of an nonholonomic mobile base equipped with a pan-tilt unit on which is mounted a camera. 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 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 task thanks to visual servoing techniques [3].

We have chosen to separately design the necessary controllers to dissociate at best the different degrees of freedom of the camera. Analyzing the robot structure shows that the two control inputs ω_r and ω_p appear to be redundant. Although this property can be used to perform additional objectives such as obstacle avoidance[5] for example , here, we have simply chosen to fix ω_p to zero, so that we control the horizontal position of the features in the image using a unique controller. Another interest in using ω_r instead of ω_p is that it allows to orientate the whole robotic system (and not only the camera) towards the targeted person, improving the task execution. Therefore, we will finally have to design three control laws to compute $(v_r, \omega_r, \omega_t)$.

[4]In our experiments, $g(.)$ is a uniform distribution over the state space.

[5]We have elaborated a first simple strategy allowing to follow a person. The obstacle avoidance problem is only roughly treated by stopping the robot when the collision risk is too high. It will be more deeply addressed through further works.

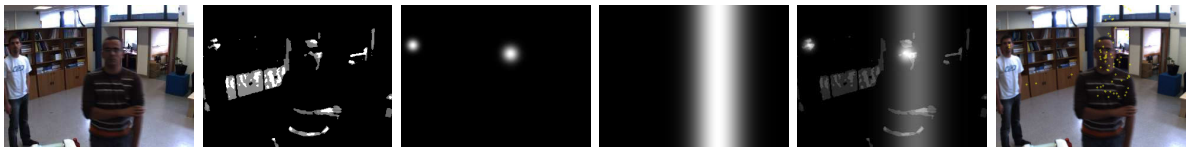


Fig. 5. From left to right: original image, skin probability image4, face detection (5), azimuthal angle from RFID detection (6), unified importance function (2) (without dynamic), accepted particles (yellow dots) after rejection sampling.

B. Control design

We have then developed three classical controllers allowing to orientate the camera and to move the robot, so that the person to be followed is always kept in its line of sight. To this aim, the idea is to exploit the tracker information. Indeed, from the images provided by the camera, this tracker can characterize the coordinates of the head gravity center in the image (u_{gc}, v_{gc}) and its associated scale s_{gc}. Our goal is to design three vision-based PID controllers to minimize E_{ptv}. E_u represents the abscissa error in the image and can then be regulated to zero by acting on the robot angular speed ω_r . Following the same reasoning, E_v corresponds to the ordinate error in the image and can be decreased to zero thanks to the tilt velocity ω_t of the PTU and E_s corresponds to the scale error regulated to zero thanks to the robot linear velocity v_r. The proposed control laws are given as follows:

$$\left\{ \begin{array}{rcl} \omega_r & = & K_{pp}E_u + K_{ip}\int E_u dt + K_{dp}\frac{dE_u}{dt} \\ \omega_t & = & K_{pt}E_v + K_{it}\int E_v dt + K_{dt}\frac{dE_v}{dt} \\ v_t & = & K_{pv}E_s + K_{iv}\int E_s dt + K_{dv}\frac{dE_s}{dt} \end{array} \right. \quad (7)$$

where (K_{pp}, K_{ip}, K_{dp}) (respectively, (K_{pt}, K_{it}, K_{dt}) and (K_{pv}, K_{iv}, K_{dv})) are the control gains which are experimentally tuned.

However, these control laws can be used only when the target lies in the image. When the latter is lost, they cannot be applied anymore and the control strategy has to be enhanced. To this aim, we use the information provided by the RFID system. Indeed, as previously mentioned, the targeted person wears a RFID tag so that, when he/she leaves the camera field of view, it is possible to get the distance d_{tag} and the orientation θ_{tag} of the tag in the robot frame. Therefore, the direction towards which the camera must be oriented to retrieve the user is known as well as the approximated distance to the user. Our idea is then to make the camera turn until the robot faces the tag, so that the tracker can retrieve the user if possible. To this aim, we simply impose a constant value ω_r^0 for the robot angular velocity ω_r to make it turn towards this direction. The tilt velocity is controlled so that the corresponding angle is brought back to its reference position, that is the position reached after each initialization of the pan/tilt unit. With this procedure, we maximize the possibility for the tracker to detect the user again. When this event occurs, the control strategy switches back to the three vision-based controllers given above. We also impose a linear velocity whose value depends on the distance d_{tag} and on the angle θ. In this way, we try[6] to keep on satisfying the constraint on the social distance, despite the visual information loss. The robot is then kept in a close neighborhood of the user in order to ease the visual signal recovery.

V. Integration and Live Experiments

A. Rackham description and software architecture

[6]The distance d_{tag} provided by the RFID system is rather inaccurate.

Fig. 6. Rackham.

Rackham is an iRobot B21r mobile platform. Its standard equipment has been extended with one digital camera mounted on a Directed Perception pan-tilt unit, one ELO touch-screen, a pair of loudspeakers, an optical fiber gyroscope, wireless Ethernet, and the previously described RFID system (figure 6). All these devices enable Rackham to act as a service robot in utilitarian public areas. It embeds robust Human Robot interaction abilities and efficient basic navigation skills.

We focus here on the software modules named `ICU` which stands for "I see you", `RFID` and `Visuserv`, which respectively encapsulate human recognition/tracking, RFID localization, and visual servoing. The data fusion extracted from `Camera` and `RFID` is performed within the `ICU` module as vision constitutes the "central" sensor. The motions of the camera and of the robot are performed within `Visuserv` which controls both actuators modules: `Platine` and `Rflex`. These modules have been implemented within in the "LAAS" architecture [1] using C/C++ interfacing scheme. The OpenCV library is used for low-level features extraction *e.g.* edge or face detection. The entire system operates at an average framerate of 6 Hz.

B. Targeted scenario

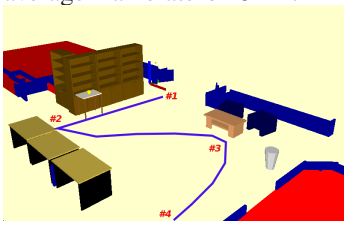


Fig. 7. Environment map showing the path of the user.

Experimental evaluations were conducted in our robotic hall in the presence of crowds. Let us recall that the goal is to make Rackham follow a non-expert tagged person in natural, dynamic, cluttered scenes in real time. Rackham is supposed to accompany the person while (i) maintaining enough space between itself and the person (to avoid collisions), (ii) respecting his/her personal space (social distance maintained within $[1.5; 2.5]$m) during the mission execution.

Figure 7 shows the environment where the experiments took place and the typical path to be followed by the RFID tagged person (see the blue line). He/she enters the hall in point $\#1$ and goes to point $\#2$ before freely walking to $\#4$ via $\#3$. Finally, he/she stops and comes back to $\#1$ to exit the hall. Thus, from $\#1$ to $\#4$, Rackham is expected to follow behind the person while from $\#4$ to $\#1$, it has to leave the person path to maintain the social distance to a suitable value. Given the overall path, the robot is expected to follow the target for about 20m. During the mission execution, other people may cross the robot trajectory, occluding the camera field of view. Both qualitative and quantitative results on trial runs are reported below in order to validate our multimodal person tracking and the proposed control strategy.

We first validate the system behavior in nominal conditions, *i.e.* without any disturbance. Note that, even in this context, the RFID tag is not always detected due to self-occlusions. Associated signals provided by the modules `ICU`, `RFID`, and `Visuserv` are collected to illustrate sequencing and synchronization aspects. Figure 9 shows a typical run of the system where the robot executes the above scenario. Figure 8 shows the corresponding evolution of these data, namely:

- two flags `ICU` and `RFID` which are respectively set to 1 when the tagged user is detected either in the image or in the RFID area located within the distance range $[0.5; 4.5]m$,
- the angle θ_{tag} and the distance d_{tag} which provides the direction where lies the tagged user with respect to the robot frame,
- the three control inputs (v_r, ω_r, ω_t) computed by the `Visuserv` module and sent to the robot by respectively `Rflex` and `Platine` modules.

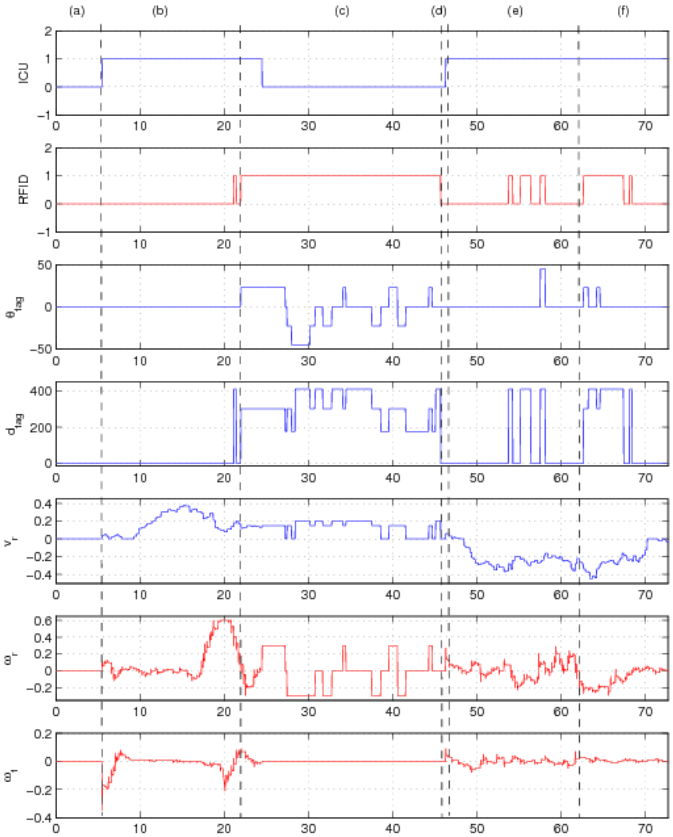


Fig. 8. Synchronization of the data flow outputs between the different modules.

Let us consider figure 9 and 8. After the mission initialization (a), the tracker focuses on the tagged person thanks to the video stream (b). The four steps of the scenario are then executed. Between points #1 and #2, the contact with the target is maintained thanks to the vision system. The control law is then computed using equations (7) on the base of the visual data provided by the tracker. The target is centered in the image while the robot adjusts its relative distance thanks to the visual template scale. Between point #2 and #4 (c), the target disappears from the camera view field (due to the joined movements of the two mobile entities), which induces a visual tracker failure. The control law is then

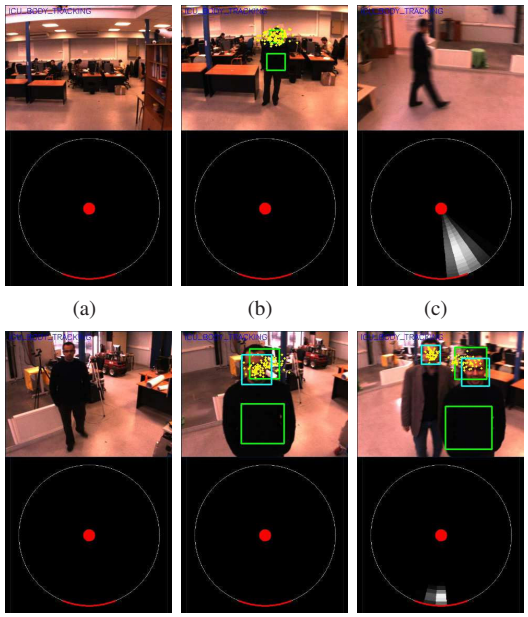


Fig. 9. Snapshots of a trial. Notice that the pan-tilt unit azimuthal position is given by the red arch on the RFID map. The blue and green squares respectively depict the face detection (user gazing the camera) and the MMSE estimate while the yellow dots represent the particles before the resampling step.

computed on the base of RFID data (θ_{tag}, d_{tag}) to make the robot face the targetted person, and converge towards him/her until d_{tag} reaches a close neighborhood of the chosen social distance value. Therefore, the person following task is still executed despite the visual target loss, while the RFID system triggers the camera in order to recover the target in the view field (d). Face detection/recognition allows to re-initialize the visual tracker while the person goes back to #1 (e). During this last part of the mission, the trajectory of the robot crosses the user's one. The robot then moves backwards to fulfill the requirement about the imposed social distance to the user. We have also quantitatively evaluated the above scenario when numerous persons lie in the robot neighborhood. This study has been performed during an informal gathering by several participants. Few of them had an advanced knowledge of our multimodal person tracker. The person following system was tested for 10 trial runs for a given targeted person. We have progressively added the number of people in the robot vicinity in order to disturb the scenario execution by inducing sporadic occlusions of the tagged person. Figure 10 shows snapshots of a typical run while the associated video and additional runs can be found at the URL `www.laas.fr/~tgerma/IROS`.

For each run, we have computed the Visual Contact Rate. This rate correspond to the ratio of the frames where the user was in the field of view over the total number of frames. This parameter indirectly measures the system robustness to artifacts such as occlusions and sporadic target losses. Table I shows these rates for the vision system only and its multimodal counterpart when increasing the number of passers-by during the scenario execution.

We can notice that the average Visual Contact Rate decreases with the increasing number of persons lying in its neighborhood while it remains almost constant for the multimodal system. These results highlights the multimodal tracker efficiency as the RFID system allows to keep the

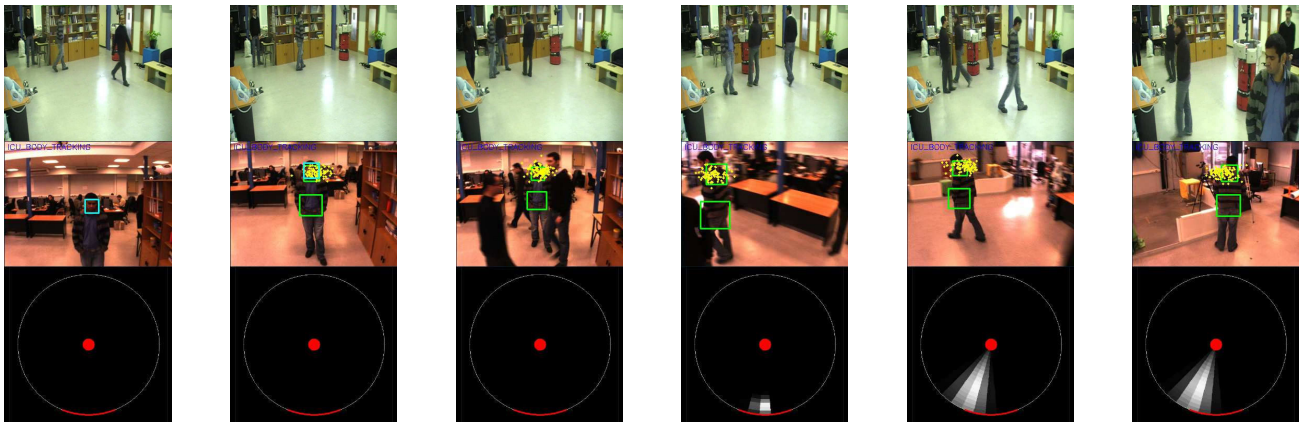


Fig. 10. Snapshots of a run for the quantitative evaluation. The first line shows the current human robot situation, the second line depicts the tracking outputs and the third line is the current RFID saliency map.

Sensor system	Number of passers-by:				Total	
	1	2	3	4	Mean	Std.
Vision only	0.22	0.50	0.47	0.19	0.35	0.23
Vision + RFID	**0.93**	**0.93**	**0.90**	**0.85**	**0.90**	**0.08**

TABLE I

RESULTS OF THE *Visual Contact Rate* WHEN CONSIDERING 1 TO 4 PASSERS-BY.

target in the visual view-field for 85% of the video stream despite the presence of four passers-by.

Finally, after the mission execution, the participants were asked to evaluate the robot's behaviour *i.e.* whether the behaviour met their expectations, how natural the behaviour was, and how appropriate the robot's following and stopping distances were. Two cases of discontent were mentioned:

1) **the robot stayed too close from the tagged person:** this situation mainly happens because of the RFID rough evaluation of the distance to the target. Sometimes, the tracker initializes to a too small scale, making the robot confuse about its distance.
2) **the robot stayed too far away from the person:** this situation is mainly due to the fact that the robot remains static when it loses both visual and RFID contacts with the user. This phenomenon appears in over-crowded scenes. As soon as RFID or face detection occurs, the tracker recovers the user and the robot continue its mission.

VI. CONCLUSION

Tracking provides important capabilities for human robot interaction and assistance of humans in utilitarian populated spaces. The paper exhibits several contributions. First, we have adapted an off-the-shelf RFID system to detect tags in an 360° view field thanks to the multiplexing of 8 antennas and embed such a system on our mobile robot Rackham to evaluate this new ID-sensor. Then, we have developed a person tracker that combines the accuracy benefits of monocular active vision with the identification certainty of such RFID-based sensor. Our technique uses the ICONDENSATION scheme, heterogeneous data driven proposals and rejection sampling mechanism to (re)-concentrate the particles on the right person during the sampling step. To our best knowledge, such data driven approach is unique in the robotics or vision literature. Finally, we have demonstrated the tracker robustness to sporadic occlusions, camera out field-of-view, and appearance changes during live experiments. These experiments have also shown that the person following task inherits the advantages of both sensor types, thereby being able to robustly track people and estimate their identity.

Several directions are currently studied regarding the whole system. Further investigations will concern the design of more compact antennas as the embeddability is essential for autonomous robots. We will also extend our tracker to multiple person as several RFID can be detected at the same time while an obstacle detection devoted to over-crowded spaces will be then outlined.

ACKNOWLEDGMENT

The authors are very grateful to Léo Bernard and Antoine Roguez for their involvement in this work which was partially conducted within the EU STREP Project Commrob funded by the European Commission Division FP6 under Contract $FP6-045441$.

REFERENCES

[1] R. Alami, R. Chatila, S. Fleury, and F. Ingrand. An architecture for autonomy. *Int. Journal of Robotic Research (IJRR'98)*, 17(4):315–337, 1998.
[2] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *Trans. on Signal Processing*, 50(2):174–188, 2002.
[3] P.I. Corke. *Visual control of robots : High performance visual servoing*. Research Studies Press LTD, 1996.
[4] T. Germa, M. Devy, R. Rioux, and F. Lerasle. A tuning strategy for face recognition in robotic application. In *Int. Conf. on Computer Vision Theory and App. (VISAPP'09)*, Lisbon, Portugal, Feb 2009.
[5] T. Germa, F. Lerasle, P. Danès, and L. Brèthes. Human/robot visual interaction for a tour-guide robot. In *Int. Conf. on Intelligent Robots and Systems (IROS'07)*, San Diego, USA, Nov 2007.
[6] D. Hahnel, W. Burgard, D. Fox, K. Fishkin, and M. Philipose. Mapping and localization with RFID technology. *Int. Conf. on Robotics and Automation (ICRA'04)*, Apr 2004.
[7] M. Isard and A. Blake. I-CONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. In *Euro. Conf. on Computer Vision (ECCV'98)*, Freibourg, Ger., Jun 1998.
[8] T. Kanda, M. Shiomi, L. Perrin, T. Nomura, H. Ishiguro, and N. Hagita. Analysis of people trajectories with ubiquitous sensors in a science museum. In *Int. Conf. on Robotics and Automation (ICRA'07)*, Roma, Italy, April 2007.
[9] J. Lee, W. Lee, and D. Jeong. Object tracking method using back-projection of multiple color histogram models. *Int. Symp. on Circuits and Systems (ISCAS'03)*, June 2003.
[10] P. Pérez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *IEEE*, 92(3):495–513, 2004.
[11] D. Schulz, D. Fox, and J. Hightower. People tracking with anonymous and ID-sensors using rao-blackwellised particle filters. In *Int. Joint Conf. on Artificial Intelligence (IJCAI'03)*, Acapulco, Mexico, August 2003.
[12] P. Viola and M. Jones. Fast multi-view face detection. In *Int Conf. on Computer Vision and Pattern Recognition (CVPR'03)*, Madison, USA, June 2003.
[13] Z. Zivkovic and B. Kröse. Part based people detection using 2D range data and images.