

Towards an Interactive Humanoid Companion with Visual Tracking Modalities

Paulo Menezes¹, Frédéric Lerasle^{2,3}, Jorge Dias¹ and Thierry Germa²
¹Institute of Systems and Robotics - University of Coimbra, ²LAAS-CNRS, ³Université Paul Sabatier
¹Portugal, ^{2,3}France

1. Introduction and framework

The idea of robots acting as human companions is not a particularly new or original one. Since the notion of “robot” was created, the idea of robots replacing humans in dangerous, dirty and dull activities has been inseparably tied with the fantasy of human-like robots being friends and existing side by side with humans. In 1989, Engelberger (Engelberger, 1989) introduced the idea of having robots serving humans in everyday environments. Since then, a considerable number of mature robotic systems have been implemented which claim to be servants, personal assistants (see a survey in Fong et al., 2003). The autonomy of such robots is fully oriented towards navigation in human environments and/or human-robot interaction.

Interaction is facilitated if the robot behaviour is as natural as possible. Two aspects of this are important. The first is to facilitate tasks, which involve direct physical cooperation between humans and robots. The second issue is that robot independent movements must appear familiar and predictable to humans. Furthermore, in order to be more effective towards a seemingly interaction, a similar appearance to humans is an important requirement. These considerations initiated probably the design of humanoid robots. One can mention here commercial robots like QRIO by Sony as well as prototypes like Alpha (Bennewitz et al., 2005), Robox (Siegwart et al., 2003), Minerva (Thrun et al., 2000) or Mobot (Nourbakhsh et al., 2003).

These systems addressed various aspects of human-robot interaction designed by a programmer. This includes all or parts of situation understanding, recognition of the human partner, understanding his intention, and coordination of motion and action and multi-modal communication. Such systems are able to communicate with its a non-expert user in a human-friendly intuitive way by employing the bandwidth of human communication and interaction modalities, typically through H/R interfaces, speech or gestures recognition. It is an evident fact that gestures are natural and rich means, which humans employ to communicate with each other, especially valuable in environments where the speech-based communication may be garbled or drowned out. Communicative gestures can represent either acts or symbols. This includes typically gestures recognition for interaction between humans and robots e.g. waving hands for good-bye, acting hello, and gestures recognition for directions to humanoid e.g. pointing out, stop motion. Unfortunately, a few of the

designed robotic systems exhibit elementary capabilities of gesture-based interaction and future developments in the robotic community will be undoubtedly devoted to satisfy this need.

Besides communication process, another and potentially deeper issue is the flexibility as humanoid robots are expected to evolve in dynamic and various environments populated with human beings. Most of the designed robotic systems lack learning representations and the interaction is often restricted to what the designer has programmed.

Unfortunately, it seems impossible to create a humanoid robot with built-in knowledge of all possible states and actions suited to the encountered situations. To face this problem, a promising line of investigation is to conceptualize cognitive robots i.e. permanent learners, which are able to evolve and grow their capacities in close interaction with non-expert users in an open-ended fashion. Some recent platforms e.g. Biron (Maas et al., 2006) or Cog (Fitzpatrick et al., 2003) enjoy these capabilities.

They have no completion and continue to learn as they face new interaction situations both with their environments and the other agents. Basically, they discover a human centred environment and build up an understanding of it. Typically, the robot companion follows a human master in his/her private home so as to familiarise it with its habitat. This human master points out specific locations, objects and artefacts that she/he believes are necessary for the robot to remember. Once such a robot has learnt, all this information, it can start interacting with its environment autonomously, for instance to share/exchange objects with humans.

The robot must also learn new tasks and actions relatively to humans by observing and try to imitate them to execute the same task. Imitation learning (Asfour, 2006), (Shon et al., 2005) addresses both issues of human-like motion and easy teaching of new tasks: it facilitates teaching a robot new tasks by a human master and at the same time makes the robot move like a human. This human instructor must have been logically beforehand identified among all the possible robot tutors, and just then granted the right to teach the robot. Activities and gestures imitation (Asfour, 2006; Nakazawa et al., 2002) is logically an essential important component in these approaches.

These reminders stress that activities/gestures interpretation and imitation, object exchange and person following are essential for a humanoid companion. Recall that two generally sequential tasks are involved in the gestures interpretation, namely the tracking and recognition stages while gestures imitation learning proceeds also through two stages: tracking, and reproduction. All these human-robot interaction modalities require, as expected, advanced tracking functionalities and impose constraints on their accuracies, or on the focus of interest. Thus, person following task requires coarse tracking of the whole human body and image-based trackers is appropriate in such situation. These trackers provide coarse tracking granularity but are generally fast and robust. Tracking hands in image plane is also sufficient to interpret many symbolic gestures e.g. a "hello" sign. On the other side, tracking hands when performing manipulation tasks requires high accuracy and so 3D-based trackers. More globally, many tasks concerning manipulation but also interaction rely on tracking of the whole upper human body limbs, and require inferring 3D information.

From these considerations, the remainder of the paper reports both on 2D and 3D tracking of the upper human body parts or hands from a single camera mounted on mobile robot as most of humanoid robots embed such exteroceptive sensor. This set of trackers is expected to fulfil the requirements of most of the aforementioned human-robot interaction modalities.

Tracking human limbs from a mobile platform has to be able to cope with: (i) automatic initialization and re-initialization after target loss or occlusions, (ii) dynamic and cluttered environments encountered by the robot during its displacements.

The paper is organized as follows. Section 2 gives a brief state-of-art related to human body parts tracking based on one or multiple cameras. This allows to introduce our approach and to highlight particle filters in our context. Our guiding principle to design both 2D and 3D trackers devoted to mobile platform is also introduced. Section 3 sums up the well-known particle filtering formalism and describes some variants which enable data fusion in this framework. The latter involve visual cues which are described in section 4. Sections 5 and 6 detail our strategies dedicated to the 2D and 3D tracking of human hands and their generalization to the whole upper human body parts. Section 7 presents a key-scenario and outlines the visual functions depicted in this paper i.e. trackers of human limbs and face recognition as trackers are classically launched as soon as the current user is identified as human master. These visual functions are expected to endow a universal humanoid robot and to enable it to act as a companion.

Considerations about the overall architecture, implementation and integration in progress on two platforms are also presented. This concerns: (i) person recognition and his/her coarse tracking from a mobile platform equipped with an arm to exchange objects with humans, (ii) fine gestures tracking and imitation by a HRP2 model as a real platform which is recently available at LAAS. Last, section 8 summarizes our contribution and opens the discussion for future extensions.

2. Related works on human body parts tracking

The literature proposes a plethora of approaches dedicated to the tracking of human body parts. Related works can be effectively organized into two broad categories, 2D or image-based tracking, and 3D tracking or motion capture. These categories are outlined in the two next subsections with special emphasis on particle filtering based approaches. Recall that activities/gestures tracking is currently coupled with recognition. Though a state of art related to activities/gestures recognition goes outside the scope of this paper, the interested reader is referred to the comprehensive surveys (Pavlovic, et al., 1997; Wu et al., 1999).

2.1 2D or image-based tracking

Many 2D tracking paradigms of the human body parts have been proposed in the literature which we shall not attempt to review here exhaustively. The reader is referred to (Gavrila, 1999; Eachter et al., 1999) for details. One can mention Kalman filtering (Schwerdt et al., 2000), the mean-shift technique (Comaniciu et al., 2003) or its variant (Chen et al., 2001), tree-based filtering (Thayanathan et al., 2003) among many others. Beside these approaches, one of the most successful paradigms, focused in this paper, undoubtedly concerns sequential Monte Carlo simulation methods, also known as particle filters (Doucet et al., 2000).

Particle filters represent the posterior distribution by a set of samples, or particles, with associated importance weights. This weighted particles set is first drawn from the state vector initial probability distribution, and is then updated over time taking into account the measurements and a prior knowledge on the system dynamics and observation models.

In the Computer Vision community, the formalism has been pioneered in the seminal paper by Isard and Blake (Isard et al., 1998a), which coins the term CONDENSATION for

conditional density propagation. In this scheme, the particles are drawn from the dynamics and weighted by their likelihood w.r.t. the measurement. CONDENSATION is shown to outperform Kalman filter in the presence of background clutter.

Following the CONDENSATION algorithm, various improvements and extensions have been proposed for visual tracking. Isard et al. in (Isard et al., 1998c) introduce a mixed-state CONDENSATION tracker in order to perform multiple model tracking. The same authors propose in (Isard et al., 1998b) another extension, named ICONDENSATION, which has introduced for the first time importance sampling in visual tracking. It constitutes a mathematically principled way of directing search, combining the dynamics and measurements. So, the tracker can take advantage of the distinct qualities of the information sources and re-initialize automatically when temporary failures occur. Particle filtering with history sampling is proposed as a variant in (Torma et al., 2003). Rui and Chen in (Rui et al., 2001) introduce the Unscented Particle Filter (UPF) into audio and visual tracking. The UPF uses the Unscented Kalman filter to generate proposal distributions that seamlessly integrate the current observation. Partitioned sampling, introduced by MacCormick and Isard in (MacCormick et al., 2000a), is another way of applying particle filters to tracking problems with high-dimensional configuration spaces. This algorithm is shown to be well suited to track articulated objects (MacCormick et al., 2000b). The hierarchical strategy (Pérez et al., 2004) constitutes a generalization.

2.2 3D tracking or motion capture

In the recent years, special devices such as data glove (Sturman et al. 1994), immersive environment (Kehl et al., 2004) and marker-based optical motion capturing system (generally Elite or VICON) are commonly used, in the Robotics community, to track the motion of human limbs. Let us mention some developments, which aim at analyzing raw motion data, acquired from the system VICON and reproduct them on a humanoid robot to imitate dance (Nakazawa et al., 2002) or walking gait (Shon et al., 2005). Using such systems is not intuitive and questionable in human-robot interaction session. Firstly, captured motion cannot be directly imported into a robot, as the raw data must be converted to its joint angle trajectories. Secondly, usual motion capture systems are hard to implement while using markers is restrictive.

Like many researchers of the Computer Vision community, we aim at investigating marker-less motion capturing systems, using one or more cameras. Such a system could be run using conventional cameras and without the use of special apparel or other equipment. To date, most of the existing marker-less approaches take advantage of the *a priori* knowledge about the kinematics and shape properties of the human body to make the problem tractable. Tracking is also well supported by the use of 3D articulated models which can be either deformable (Heap et al., 1996; Lerasle et al., 1999; Kakadiaris et al., 2000; Metaxas et al., 2003; Sminchisescu et al., 2003) or rigid (Delamarre et al., 2001; Giebel et al., 2004; Stenger et al., 2003). In fact, there is a trade-off between the modelling error, due to rigid structures, the number of parameters involved in the model, the required precision, and the expected computational cost. In our case, the creation of a simple and light approach that would be adequate to for a quasi-real-time application was one of the ideas that guided the developments. This motivated our choice of using truncated rigid quadrics to represent the limbs' shapes. Quadrics are, indeed, quite popular geometric primitives for use in human body tracking (Deutcher et al., 2000; Delamarre et al., 2001; Stenger et al., 2003). This is due

to the fact that they are easily handled, and can be combined to create complex shapes, and their projections are conic sections that can be obtained in closed form. Our projection method that will be depicted later, although being inspired from (Stenger et al., 2001) has the advantage that it requires less computational power than this one.

Two main classes of 3D model-based trackers can be considered, 3D reconstruction-based approaches (Delamarre et al., 2001; Urtasun et al., 2004) and appearance-based approaches, being both widely investigated. While the former performs a reconstruction of the largest possible number of points of the tracked object or structure and then tries to match them in 3D space, the latter tries to solve the problem of in which configuration should the target be for its representation being the currently observed one. Normally some characteristic features of the object are used to in the construction of a model-to-image fitting process. Our work that is presented in this paper is focused on the use of this kind of approach making no assumptions about clothing and background structure.

To cope with the lack of discriminant visual features, the presence of clutter, and the frequent occurrence of mutual occlusions between limbs, one solution is to base the observation model on multiple views (Delamarre et al., 2001; Deutscher et al., 2000; Gavrila et al., 1996; Lerasle et al., 1999; Stenger et al., 2001; Urtasun et al., 2004). Another solution (Gonçalves et al., 1995; Park et al. 2003; Sidenbladh et al., 2000; Sminchisescu et al., 2003), which is the one we have chosen, is to use a single view and increase the reliability and specificity of the observation model. To do so, a robust and probabilistically motivated integration of multiple measurement modalities is of great help. There are several examples in the literature of such integration like, for example edges and colour cues in (Stenger et al., 2003), edges/silhouette and motion cues in (Sminchisescu et al., 2003) or edges, texture and 3D data cues in (Giebel et al., 2004). In our case, we propose an observation model that combines edges and motion cues for the quadrics limbs, with local colour and texture patches on clothing acting as natural markers. Finally and inspired from (Sminchisescu et al., 2003), we add joints limits and self-body collision removal constraints to the overall model.

Regarding the tracked movements, some approaches rely on simplifications brought in either by using sophisticated learnt motion models, such as walking (Urtasun et al., 2004), or by restricting movements to those contained roughly in a fronto-parallel plane (Sidenbladh et al., 2000). Both simplification choices are well suited to monocular approaches. No specific motion models are used in this work, as we want to be able to track general human motions. In such unconstrained setup, a monocular estimation process suffers necessarily from the inevitable multi-modality of the observation process.

Each of these solutions produces a local minimum in the observation function, by consequence when any single-hypothesis-tracker is started in a position of configuration space too far from the good one, it may simply be trapped in one of the false minima, with the consequent tracking failure target loss.

Reliable tracking requires a powerful multiple hypothesis tracker capable of finding and following a significant number of minima. Local descent search strategies (Delamarre et al., 2001; Lerasle et al., 1999; Kakadiaris et al., 2000; Rehg et al., 1995; Urtasun et al., 2004) do search a local minimum, but with multi-modality there is no guaranty that the globally most representative one is found. Like others (Deutscher et al., 2000; Poon et al., 2002; Wu et al., 2001), we address these problems by employing particle-filtering techniques for the following reasons.

Particle filtering generates random sampling points according to a proposal distribution, which may contain multiple modes encoding "the good places to look at". Such probabilistic framework allows the information from different measurements sources to be fused in a principled manner. Although this fact has been acknowledged before, it has not been fully exploited for 3D trackers. Combining a host of cues such as colour, shape, and even motion, may increase the reliability of estimators dedicated to track human limbs.

In what concerns the computational cost, particle filters techniques normally require a substantial computation power, especially in high state-space dimensionality cases, which make the number of required samples to explode. Consequently, large efforts have been devoted to tackle such problem by reducing both the model's dimension through PCA (Wu et al., 2001; Uratasum et al., 2004), and the number of samples by testing stochastic sampling "variants" (Deutscher et al., 2000; Sminchisescu et al., 2003).

2.3 Problem statement and guiding principle

2D or 3D human tracking from a mobile platform is a very challenging task, which imposes several requirements. First, the sensor's setup, is naturally embedded on the autonomous robot. By consequence from the camera point of view all scene objects move, this precludes the use of some useful techniques like background subtracting for isolating the target objects. As the robot is expected to evolve in environments that are highly dynamic, cluttered, and frequently subjected to illumination changes, several hypotheses must be handled simultaneously by the trackers. This is due to the multi-modality that appears in the statistical distributions of the measured parameters, as a consequence of the clutter or the changes in the appearance of the target. Consequently, several hypotheses must be handled simultaneously in the developed trackers, and a robust integration of multiple visual cues is required to efficiently localize the good likelihood peaks. Finally, on-board computational power is limited so that only a small percentage of these resources can be allocated to tracking, the remaining part being required to enable the concurrent execution of other functions as well as decisional routines within the robot's architecture. Thus, care must be taken to design efficient algorithms.

The particle-filtering framework is well suited to the above requirements and is widely used in the literature both for 2D or 3D tracking purpose. The popularity of this framework is due to its simplicity, ease of implementation, and modelling flexibility. This framework makes no restrictive assumptions about the probability distributions and enables the fusion of diverse measurements in a simple way. Clearly, combining a host of cues may increase our trackers versatility and reliability. Finally, from the numerous particle-filtering strategies proposed in the literature, one is expected to fit to the requirements of each tracker modality. These considerations lead us to investigate on particle filtering strategies for data fusion. The creation of simple and light monocular-based trackers that would adequate to for a quasi-real time application was another motivation that guided our developments.

3. Particle filtering algorithms for data fusion

3.1 Generic algorithm

Particle filters are sequential Monte Carlo simulation methods for the state vector estimation of any Markovian dynamic system subject to possibly non-Gaussian random inputs (Arulampalam et al., 2002). The aim is to recursively approximate the posterior density

function (pdf) of the state vector x_k at time k conditioned on the set of measurements $z_{1:k} = z_1, \dots, z_k$ through the linear point-mass combination

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

which expresses the selection of a value -or "particle"- $x_k^{(i)}$ with probability -or "weight"- $w_k^{(i)}, i = 1, \dots, N$.

A generic particle filter or SIR is shown on Table 1. The particles $x_k^{(i)}$ evolve stochastically over the time, being sampled from an importance density $q(\cdot)$, which aims at adaptively exploring "relevant" areas of the state space. Their weights $w_k^{(i)}$ are updated thanks to $p(x_k^{(i)} | x_{k-1}^{(i)})$ and $p(z_k | x_k^{(i)})$, respectively the state dynamics and measurement functions, so as to guarantee the consistency of the approximation (1). In order to limit the degeneracy phenomenon, which says that after few instants the weights of all but one particle tend to zero, step 8 inserts a resampling process. Another solution to limit this effect in addition to re-sampling is the choice of a good importance density.

$\{\{x_k^{(i)}, w_k^{(i)}\}_{i=1}^N\}$ = SIR PF($\{\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}_{i=1}^N, z_k\}$)
1. IF $k = 0$, THEN Draw $x_0^{(1)}, \dots, x_0^{(i)}, \dots, x_0^{(N)}$ i.i.d. according to $p(x_0)$, and set $w_0^{(i)} = \frac{1}{N}$ END IF
2. IF $k \geq 1$ THEN $\{\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}_{i=1}^N\}$ being a particle description of $p(x_{k-1} z_{1:k-1})$ →
3. FOR $i = 1, \dots, N$, DO
4. "Propagate" the particle $x_{k-1}^{(i)}$ by independently sampling $x_k^{(i)} \sim q(x_k x_{k-1}^{(i)}, z_k)$
5. Update the weight $w_k^{(i)}$ associated to $x_k^{(i)}$ according to the formula $w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(z_k x_k^{(i)}) p(x_k^{(i)} x_{k-1}^{(i)})}{q(x_k^{(i)} x_{k-1}^{(i)}, z_k)}$, prior to a normalization step so that $\sum_i w_k^{(i)} = 1$
6. END FOR
7. Compute the conditional mean of any function of x_k , e.g. the MMSE estimate $E_{p(x_k z_{1:k})}[x_k]$, from the approximation $\sum_{i=1}^N w_k^{(i)} \delta(x_k - x_k^{(i)})$ of the posterior $p(x_k z_{1:k})$
8. At any time or depending on an "efficiency" criterion, resample the description $\{\{x_k^{(i)}, w_k^{(i)}\}_{i=1}^N\}$ of $p(x_k z_{1:k})$ 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 $x_k^{(i)}$ and $w_k^{(i)}$ with $x_k^{(s^{(i)})}$ and $\frac{1}{N}$
9. END IF

Table 1. Generic particle filtering algorithm (SIR)

3.2 Importance sampling from either dynamics or measurements: basic strategies

The CONDENSATION algorithm is instanced from the SIR algorithm as $q(x_k | x_{k-1}, z_k) = p(x_k^{(i)} | x_{k-1}^{(i)})$. A difference relative to the SIR algorithm is that the re-sampling step 8 is applied on every cycle. Resampling by itself cannot efficiently limit the degeneracy phenomenon as the state-space is blindly explored without any knowledge of the observations. On the other side, the ICONDENSATION algorithm (Isard et al., 1998), considers an importance density $q(\cdot)$, which classically relates to the importance function $\pi(x_k^{(i)} | z_k)$ defined from the current image. However, if a particle drawn exclusively from the image is inconsistent with its predecessor in terms of state dynamics, the update formula

leads to a small weight. An alternative consists in sampling the particles according to the measurements, dynamics and the prior, so that, with $\alpha, \beta \in [0; 1]$

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

3.3 Towards the “optimal” case: the Auxiliary Particle Filter

The Auxiliary Particle Filter (Pitt et al., 1999) noted APF depicted by the algorithm of Table 2 is another variant that aims to overcome some limitations of the “blind exploration”. This algorithm considers an auxiliary density $p(z_k | \mu_k^{(i)})$, where $\mu_k^{(i)}$ characterise the density of x_k conditioned on $x_{k-1}^{(i)}$ (step 4). Compared to the CONDENSATION scheme, the advantage of this filter is that it naturally generates points from the sample at $k-1$ which, conditioned on the current measure, are most likely to be close to the true state and so improve the estimate accuracy. In practice, it runs slightly slower than the CONDENSATION as we need to evaluate the auxiliary weights $\lambda_k^{(i)}$ (step 4) and to perform two weighted bootstraps (steps 4 and 9) rather than one. However, the improvement in sampling will usually dominate these small effects. By making proposals that have high conditional likelihoods, we reduce the cost of sampling many times from particles, which have very low likelihoods and so will not be re-sampled at the second process stage. This improves the statistical efficiency of the sampling procedure and it means that we can reduce substantially the number N of particles.

$\{x_k^{(i)}, w_k^{(i)}\}_{i=1}^N = \text{AUXILIARY PF}(\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}_{i=1}^N, z_k)$	
1:	IF $k = 0$, THEN Draw $x_0^{(1)}, \dots, x_0^{(i)}, \dots, x_0^{(N)}$ i.i.d. according to $p(x_0)$, and set $w_0^{(i)} = \frac{1}{N}$ END IF
2:	IF $k \geq 1$ THEN $\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}_{i=1}^N$ being a particle description of $p(x_{k-1} z_{1:k-1})$ —
3:	FOR $i = 1, \dots, N$, DO
4:	From the approximation $\hat{p}(z_k x_{k-1}^{(i)}) = p(z_k \mu_k^{(i)})$ —e.g. with $\mu_k^{(i)} \sim p(x_k x_{k-1}^{(i)})$ or $\mu_k^{(i)} = E_{p(x_k x_{k-1}^{(i)})}[x_k]$ —, compute the auxiliary weights $\lambda_k^{(i)} \propto w_{k-1}^{(i)} \hat{p}(z_k x_{k-1}^{(i)})$, prior to a normalization step so that $\sum_i \lambda_k^{(i)} = 1$
5:	END FOR
6:	Resample $\{x_{k-1}^{(i)}, \lambda_k^{(i)}\}_{i=1}^N$ —or, equivalently, sample in $\{1, \dots, N\}$ the indexes $s^{(1)}, \dots, s^{(N)}$ of the particles at time $k-1$ according to $P(s^{(i)} = j) = \lambda_k^{(j)}$ —in order to get the equivalent evenly weighted particles set $\{x_{k-1}^{(i)}, \frac{1}{N}\}_{i=1}^N$; both $\sum_{i=1}^N \lambda_k^{(i)} \delta(x_{k-1} - x_{k-1}^{(i)})$ and $\frac{1}{N} \sum_{i=1}^N \delta(x_{k-1} - x_{k-1}^{(i)})$ approximate the smoothing pdf $p(x_{k-1} z_{1:k})$
7:	FOR $i = 1, \dots, N$, DO
8:	“Propagate” the particles by independently drawing $x_k^{(i)} \sim p(x_k x_{k-1}^{(i)})$
9:	Update the weights, prior to their normalization, by setting $w_k^{(i)} \propto \frac{p(z_k x_k^{(i)})}{\hat{p}(z_k x_{k-1}^{(i)})} = \frac{p(z_k x_k^{(i)})}{p(z_k x_{k-1}^{(i)})}$
10:	Compute the conditional mean of any function of x_k , e.g. the MMSE estimate $E_{p(x_k z_{1:k})}[x_k]$, 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:	END FOR
12:	END IF

Table 2. Auxiliary Particle Filter (APF)

4. Importance and measurement functions

Importance functions $\pi(\cdot)$ involve generally discriminant but possibly intermittent visual cues while measurement functions $p(z|x)$ involve cues which must be persistent but are however more prone to ambiguity for cluttered scene (Pérez et al., 2004).

Combining or fusing multiple cues enables the tracker to better benefit from distinct information, and confers robustness w.r.t temporary failures. Measurement and importance functions are depicted hereafter as well as some considerations regarding data fusion.

4.1 Measurement functions

4.1.1 Shape cue

The use of shape cues requires indeed that the class of targets to be tracked is known *a priori* and that contour models can be learnt beforehand, i.e. that coarse 2D ou 3D models of the targeted limbs can be used. For simple view-based shape representation, human limbs are therefore represented by coarse silhouette contours (Figure 1). For 3D tracking, a preliminary 3D model projection and hidden parts removal is required (Delamarre et al, 2001; Deutscher et al., 2001; Menezes et al 2005a, Sminchisescu et al., 2003).



Figure 1. Examples of silhouette templates

The shape-based likelihood is classically computed using the sum of the squared distances between model points $x(j)$ and the nearest closest edges $z(j)$, which lie on the normals that pass through the points $x(j)$. These measurement points are chosen uniformly distributed along the model.

$$p(z^S|x) \propto \exp\left(-\lambda_s \frac{D^2}{2\sigma_s^2}\right), \quad D = \frac{1}{N_p} \sum_{j=1}^{N_p} |x(j) - z(j)|. \quad (2)$$

where λ_s is a weight dedicated to further 3D tracking purpose (see section 6.2), j indexes the N_p model points, and σ_s a standard deviation being determined a priori.

A variant (Giebel et al., 2004) consists in converting the edge image into a Distance Transform image, noted I_{DT} which is used to peek the distance values. The advantage of matching our model contours against a DT image rather than using directly the edges image is twofold. Firstly, the similarity measure D is a smoother function of the model pose parameters. Secondly, this reduces the involved computations as the DT image can be generated only once, independently of the number N of particles used in the filter. The distance D becomes

$$D = \frac{1}{N_p} \sum_{j=1}^{N_p} I_{DT}(j), \quad (3)$$

where $I_{DT}(j)$ is the associated value in the DT image. Figure 1 (a) and (b) shows two plots of these two likelihoods for an image-based tracker where the target is a 2D elliptical template corresponding coarsely to the head of the right subject in the input image. As expected, the distance (3) appears to be less discriminant to clutter but is shown to enjoy least time consumption for $N \geq 100$.

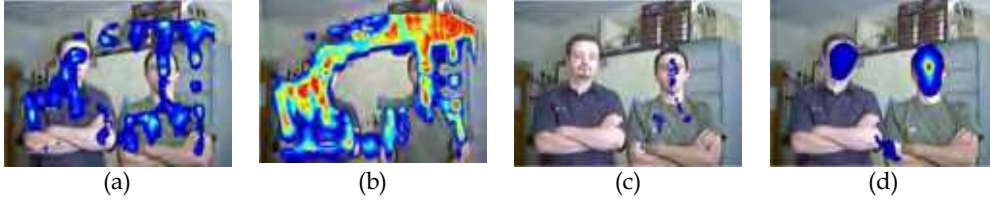


Figure 2. Likelihoods for 2D tracker for shape (a)-(b), shape and optical flow (c), shape and colour (d)

4.1.2 Shape and motion cues combination

In our context, and contrary to the background, which is assumed to remain static, the human limbs are expected to be moving, possibly intermittently. To cope with cluttered scenes and reject false background attractors, we favour the moving edges, if they exist, as they are expected to correspond to the moving target. As the target can be temporarily stopped, the static edges are not completely rejected, but only made less attractive than the moving ones. The points $z(j)$ in (2) receive the additional constraint that the corresponding optical flow vectors $\vec{T}(z(j))$ must have nonzero norm. The new likelihood $p(z^{MS}|x)$ involves the following similarity measure

$$D = \frac{1}{N_p} \sum_{j=1}^{N_p} |x(j) - z(j)| + \rho \gamma(z(j)), \quad (4)$$

where $\gamma(z(j)) = 0$ (resp. 1) if $\vec{T}(z(j)) \neq 0$ (resp. if $\vec{T}(z(j)) = 0$) and $\rho > 0$ terms a penalty. Figure 2-(c) plots this more discriminant likelihood function for the example seen above. The target is still the subject on the right, but is assumed to be moving.

Regarding the similarity measure (3), shape and motion cues are combined by using two DT images, where the second one $I'_{DT}(j)$ is obtained by filtering out the static edges, based on the local optical flow vector. The distance D becomes

$$D = \frac{1}{N_p} \sum_{j=1}^{N_p} \min(I_{DT}(j), K.I'_{DT}(j)). \quad (5)$$

where weight values $K \leq 1$ make moving edges more attractive.

4.1.3 Colour cue

Clothes colours create a clear distinction between the observed persons but also the limbs (head, hands and feet, trunk of sleeves) for a given person. Consequently, using clothing patches of characteristic colour distributions, i.e. natural markers, seems very promising. Reference colour models are associated with these targeted ROIs. For a given ROI, we denote h_{ref}^c and h_x^c two N_{bi} -bin normalized histograms in channel c corresponding respectively to the target and a region B_x related to any state x. The colour likelihood model must be defined so as to favour candidate histograms h_x^c close to the reference histogram h_{ref}^c . The likelihood has a form similar to (2), provided that D terms the Bhattacharyya distance (Pérez et al., 2004) between the two histograms. The latter can also depict the similarity of several colour patches related to faces but also clothes, each with its own

reference histogram. Let the union $B_x = \bigcup_{p=1}^{N_R} B_{p,x}$ be associated with the set of N_R reference histograms $\{h_{p,ref}^c : c \in \{R, G, B\}, p = 1, \dots, N_R\}$. By assuming conditional independence of the colour measurements, the likelihood $p(z^C|x)$ becomes

$$p(z^C|x) \propto \exp\left(-\sum_c \sum_{p=1}^{N_R} \lambda_{p,c} \frac{D^2(h_{p,x}^c, h_{p,ref}^c)}{2\sigma_c^2}\right), \quad (6)$$

where $\lambda_{p,c}$ are weighting factors dedicated to further 3D tracking purpose (see section 6.2). This multi-part extension is more accurate thus avoiding the drift, and possible subsequent loss, experienced sometimes by the single-part version (Pérez et al., 2002). Figure 2-(d) plots this likelihood function $p(z^C|x)$ for the above example. Let us note that, from (6), we can also decline a likelihood value $p(z^T|x)$ relative to textured patches based on the intensity component.

4.1.4 Multiple cues fusion

Assuming the measurement models to be mutually independent given the state. Given M measurement sources (z^1, \dots, z^M) , the global measurement function can be factorized as

$$p(z^1, \dots, z^M|x) \propto \prod_{m=1}^M p(z^m|x). \quad (7)$$

As mentioned before, data fusion is also required for 3D tracking in order to efficiently localize the good likelihood peak in the state space. Figure 3-left shows the plot of the likelihood $p(z^S|x)$ involving the distance (3) and obtained by sweeping a subspace of the configuration space formed by two parameters of a human arm 3D model. Figure 3-middle plots an approximation of the coloured multi-patches likelihood $p(z^C|x)$ entailed in our tracker. The reference colour ROI corresponds to the marked hand. Fusing shape, motion and colour, as plotted in Figure 3-right, is shown to be more discriminant as expected.

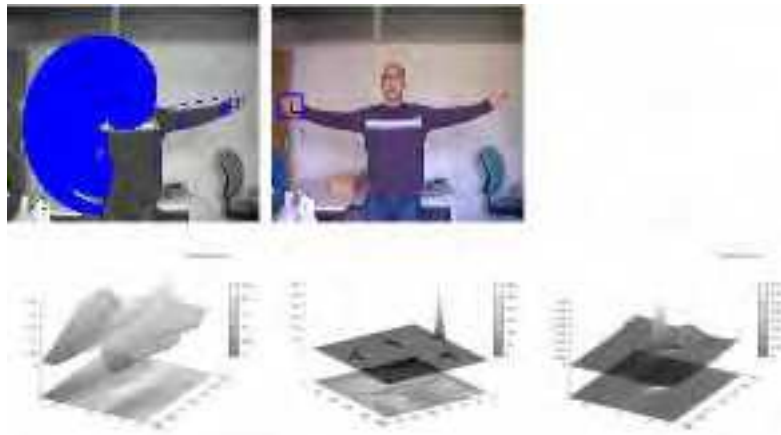


Figure 3. Likelihood plots for 3D tracking: shape cue, colour cue, both shape and colour cues

Clearly, mixing all these cues into the measurement function of the underlying estimation scheme helps our 2D or 3D trackers to work under a wide range of conditions encountered by our robot during its displacements.

4.2 Importance function

4.2.1 Shape cue

This importance function $\pi(\cdot)$ considers the outputs from face or hand detectors. Our face detection system is based on the AdaBoost algorithm and uses a boosted cascade of Haar-like features. Each feature is computed by the sum of all pixels in rectangular regions, which can be computed very efficiently using integral images. The idea is to detect the relative darkness between different regions like the region of the eyes and the cheeks (Figure 4-left).

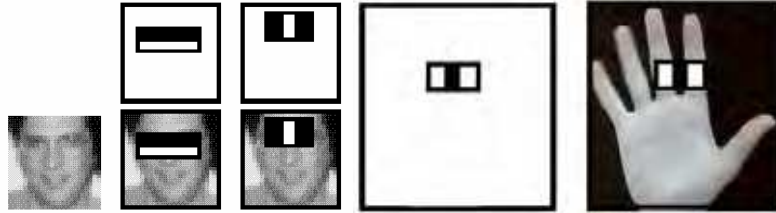


Figure 4. First Haar features for faces (Viola et al., 2001) and for hands

Originally, this idea was developed by Viola et al. in (Viola et al., 2001) to reliably detect faces, in the range of $[-45,45]$ degrees of out-of-plane rotation, without requiring a skin colour model. This widely used detector works quickly and yields high detection rates.

This idea was extended for detecting hands. Our classifier was trained with 2000 images containing upright hands, and 6000 images without hands and used as negative samples. This hand detector exhibits a detection rate slightly smaller than the previous, mainly due to the lack of discriminant contrasts in the hand. Figure 4-right shows example of Haar-like feature used in this context. A video relative to hand detection can be downloaded from the following URL <http://www.isr.uc.pt/~paulo/HRI>.

Let us characterize the associated importance functions. Given N_B detected faces or hands, and $\mathbf{p}_i = (u_i, v_i), i = 1, \dots, N_B$ the centroid coordinates of each such region. The associated importance function $\pi(\mathbf{x}|z^S)$ at location $\mathbf{x} = (u_k, v_k)$ follows, as the Gaussian mixture proposal

$$\pi(\mathbf{x}|z^S) \propto \sum_{i=1}^{N_B} \mathcal{N}(\mathbf{x}; \mathbf{p}_i, \text{diag}(\sigma_{u_i}^2, \sigma_{v_i}^2)). \quad (8)$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

4.2.2 Colour cue

Human skin colours have a specific distribution in colour space. Training images from the Compaq database (Jones et al., 1998) enables to construct the associated distributions. The detection of skin-colored blobs is performed by subsampling the input image prior to grouping the classified skin-like pixels. Parts of the segmented regions are filtered regarding their aspect ratio. Then, the importance function $\pi(\mathbf{x}|z^S)$ is defined from the resulting blobs by a Gaussian mixture similar to (8).

4.2.3 Multi-cues mixture

In a mobile robotic context, the efficiency of the above detection modules is influenced by the variability of the environment clutters and the change of viewing conditions. Therefore, the importance function $\pi(\cdot)$ can be extended to consider the outputs from any of the M detectors, i.e.

$$\pi(\mathbf{x}|z^1, \dots, z^M) = \frac{1}{M} \sum_{j=1}^M \pi(\mathbf{x}|z^j). \quad (9)$$

5. Image-based tracking dedicated to upper human body parts

5.1 Preliminary works for hands tracking

Preliminary investigations (Menezes et al., 2004c) deal with an image-based tracker suitable to estimate fronto-parallel motions of the hand e.g. when performing a "hello" or a "halt" sign. The aim is to fit the view-based template relative to the targeted hand all along the video stream, through the estimation of its image coordinates (u,v) , its scale factor s , and its orientation θ . All these parameters are accounted for in the state vector \mathbf{x}_k related to the k -th frame. With regard to the dynamics model $p(\mathbf{x}_k|\mathbf{x}_{k-1})$, the image motions of observed people are difficult to characterize over time. This weak knowledge is thus formalized by defining the state vector as $\mathbf{x}_k = (u_k, v_k, s_k, \theta_k)$ 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 $\mathcal{N}(\cdot|\mu, \Sigma)$ terms the Gaussian distribution with mean μ and covariance $\Sigma = \text{diag}(\sigma_u^2, \sigma_v^2, \sigma_s^2, \sigma_\theta^2)$.

Complex filtering strategies are not necessary in this tracker and we opt logically for the CONDENSATION algorithm as it enjoys the least time consumption. The tracker is launched automatically when detecting hands after agreement between both the Haar-like features based detector and the skin blobs detector outcomes. The particle-weighting step entails the likelihood $p(z^{MS}|\mathbf{x})$ based on the similarity measure (4) and a hand silhouette template (Figure 1). Characteristics and parameter values reported in Table 3 are used in the likelihoods, proposal and state dynamics involved in our hand tracker.

Symbol	Meaning	Value
-	Particle filtering strategy	CONDENSATION
N	Number of particles	100
(nbL, nbC)	Image resolution	(320,240)
σ_s	standard deviation in $p(z^{MS}/x_k)$	36
N_p	Number of model points in similarity measure	30
ρ	Penalty in similarity distance	0.12

Table 3. Characteristics and parameter values for our image-based hand tracker

The running time of this tracker is about 50fps on a PentiumIV-3GHz. Figure 5 shows some image-based tracking snapshots from a sequence involving heavy cluttered background. The entire video can be found at the URL <http://www.isr.uc.pt/paulo/HRI>. This elementary and specific tracker has not been integrated in the Jido's software architecture (detailed in section 7.1).



Figure 5. Image-based tracking of hand in heavy cluttered background

5.2 Extension to the upper human body limbs

Same guiding principles, namely data fusion in an appropriate particle filtering strategy, were used to develop an image-based tracker dedicated to the upper human body parts. This tracker can typically be launched for: (i) person following i.e. coordinating the robot's displacements, even if only coarsely, with those of the tracked robot user, (2) people perception in the robot vicinity, for instance to heckle them. This coarse human tracking is used to plan how to position the robot with respect to human beings in a socially acceptable way.

Unfortunately, more than one authorised person can be in robot vicinity, what could make the tracker continuously switch from the targeted person to another. Therefore, for re-identifying individuals information based on face recognition and clothing colour are logically entailed in the characterization of the tracker. These permit to distinguish individuals but also to recover the targeted person after temporary occlusions or out-of-sight. Moreover, any person must be, normally recognized among the potential human masters database before receiving the grant to learn the robot.

5.2.1 Face recognition

This function aims to classify bounding boxes \mathcal{F} of detected faces (see section 4.2) into either one class C_i out of the set $\{C_i\}_{1 \leq i \leq M}$ -- corresponding to M users faces presumably learnt offline -- or into the void class C_\emptyset . Our approach, clearly inspired by (Turk et al., 1991), consists in performing PCA, and keeps as eigenface bases the first eigenvectors accounting on a certain average of the total class variance. Our evaluations are performed on a face database that is composed of 6000 examples of $M=10$ individuals acquired by the robot in a wide range of typical conditions: illumination changes, variations in facial orientation and expression, etc. The database is separated into two disjoint sets: (i) the training set (dedicated to PCA) containing 100 images per class, (ii) the test set containing 500 images per class. Each image is cropped to a size of 30x30 pixels. To improve the method, two lines of investigations have been pursued.

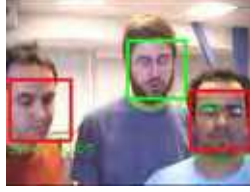


Figure 6. Example of face recognition

Firstly, evaluations highlight that the Distance-From-Face Space (DFFS) error norm leads to the best performances in term of classification rate. For a given face $\{\mathcal{F}(j), j \in \{1, \dots, 30 \times 30\}\}$, the DFFS criteria is written as follows

$$\mathcal{D} = \sum_{j=1}^{30 \times 30} |\mathcal{F}(j) - \mathcal{F}_r(j)| \quad (10)$$

where \mathcal{F}_r is the reconstructed image after projection of \mathcal{F} onto a PCA basis. For a set of M learnt tutors (classes) noted $\{C_l\}_{1 \leq l \leq M}$ and a detected face \mathcal{F} , we can define for each class C_l , the distance $\mathcal{D}_l = \mathcal{D}(\mathcal{F}, C_l)$ and an a priori probability $P(C_l|\mathcal{F})$ of labelling to C_l

$$\begin{cases} P(C_0|\mathcal{F}) = 1 \text{ and } \forall l, P(C_l|\mathcal{F}) = 0 \text{ when } \forall l, \mathcal{D}_l > \tau \\ \forall l, P(C_l|\mathcal{F}) = \frac{\mathcal{N}(\mathcal{D}_l|0, \Sigma_r)}{\sum_p \mathcal{N}(\mathcal{D}_p|0, \Sigma_r)} \text{ otherwise,} \end{cases} \quad (11)$$

where τ is a threshold predefined automatically, C_0 refers the void class and h terms the Heaviside - or "step" - function: $h(x)=1$ if $x>0$, 0 otherwise.

Secondly, from the Heseltine et al. investigations in (Heseltine et al., 2002), we evaluate and select the most meaningful image pre-processing in terms of false positives and false negatives. We plot ROC curves based on different image pre-processing techniques for our error norm \mathcal{D} . These ROC curves are shown as the sensitivity (the ratio of true positives over total true positives and false negatives) versus the false positive rate. Histogram equalization is shown to outperform the other techniques for our database (Figure 7).

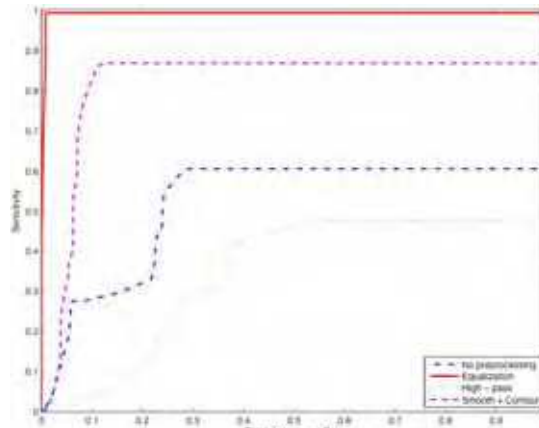


Figure 7. ROC curves image for different pre-processing techniques

Figure 6 shows a snapshot of detected (red)/recognized (green) faces with associated probabilities for a targeted person named Sylvain. More details on the face recognition process can be found in (Germa et al. 2007) and at the URL <http://www.laas.fr/~tgerma>. This face classifier relates to the module HumRec in the Jido's software architecture (see section 7.1).

5.2.2 Image-based tracking

This tracker is inspired from previously developed ones detailed in (Brèthes et al., 2005). It involves the state vector $\mathbf{x}_k = [u_k, v_k, s_k]$ - the orientation θ_k being set to a known constant. Regarding the filtering strategy, we opt for the ICONDENSATION scheme, which allows automatic initialization, and aid recovery from transient tracking failures thanks to detection modules. Let us characterize both importance and measurement functions involved in the tracker. The importance function mixes, thanks to (9) the outputs from the colour blobs and face detectors. The importance function (7) becomes

$$\pi(\mathbf{x}_k | z_k^S) \propto \sum_{i=1}^{N_B} P(C_i | \mathcal{F}_i) \mathcal{N}(\mathbf{x}; \mathbf{p}_i, \text{diag}(\sigma_{u_i}^2, \sigma_{v_i}^2)). \quad (12)$$

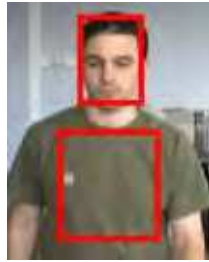


Figure 8. A two-colour patch template template

Two colour models $h_{ref_1}^c$ and $h_{ref_2}^c$ are considered in the colour-based likelihood $p(z_k^C | \mathbf{x}_k)$, respectively for the head and the torso of the guided person (Figure 19 and 10). Their initializations are achieved according to frames, which lead to $P(C_i | \mathcal{F})$ probabilities equal to one. In the tracking loop, the colour model $h_{ref_2}^c$ is re-initialized with the initial values when the user verification is highly confident, typically $P(C_i | \mathcal{F}_i) = 1$. When the appearance of these two ROIs is supposed to change in the video stream, the target reference model is updated from the computed estimates through a first-order filtering process i.e.

$$h_{ref,k}^c = (1 - \kappa) \cdot h_{ref,k-1}^c + \kappa \cdot h_{E[x_k]}^c, \quad (13)$$

where κ weights the contribution of the mean state histogram $h_{E[x_k]}^c$ to the target model $h_{ref,k-1}^c$, and index p has been omitted for compactness reasons. The models updating can lead to drifts with the consequent loss of the target. To avoid such tracker failures, the global measurement model fuses, thanks to (7), colour but also shape cues. The shape-based likelihood $p(z_k^S | \mathbf{x}_k)$ entails the similarity distance (3) and the head silhouette template (Figure 1). Characteristics and parameter values describing the likelihoods, state dynamics are listed in Table 4.

Due to the efficiency of the face recognition proposal (12), good tracking results are achieved with a reasonably small number of particles i.e. $N=150$ particles. A PentiumIV-3GHz requires about 40 fps to process the tracking.

Figure 9 and 10 show snapshots of two typical sequences in our context. All regions -- centred on the yellow dots -- close to detected faces with high recognition probabilities corresponding to the person on the background are continually explored. Those -- in blue colour -- that do not comply with the targeted person are discarded during the importance-sampling step. Recall that, for large range out-of-plane face rotations ($> |45^\circ|$), the proposal continues to generate pertinent hypotheses from the dynamic and the skin blobs detector. The green (resp. red) rectangles represent the MMSE estimate in step 7 of Table 1 with high (resp. low) confidence in the face recognition process. The proposal generates hypotheses (yellow dots) in regions of significant face recognition probabilities.

Symbol	Meaning	Value
-	Particle filtering strategy	ICONDENSATION
N	Number of particles	150
(nbL, nbC)	Image resolution	(320, 240)
(α, β)	Coeff. in importance function $q(\mathbf{x}_k \mathbf{x}_{k-1}, z_k)$	(0.3, 0.6)
$(\sigma_u, \sigma_v, \sigma_s)$	Standard deviation in random walk models	(11, 6, $\sqrt{0.1}$)
$(\sigma_{u_i}, \sigma_{v_i})$	Standard deviation in importance function $\pi(\mathbf{x}_k z^S)$	(6, 6)
$(\sigma_{u_i}, \sigma_{v_i})$	Standard deviation in importance function $\pi(\mathbf{x}_k z^C)$	(6, 6)
N_p	Number of model points in similarity measure	15
σ_s	Standard deviation in shape-based likelihood $p(z_k^S \mathbf{x}_k)$	1.5
N_R	Number of patches in $p(z_k^C \mathbf{x}_k)$	2
σ_c	Standard deviation in color based likelihood $p(z_k^C \mathbf{x}_k)$	0.03
N_{bi}	Number of colour bins per channel involved in $p(z_k^C \mathbf{x}_k)$	32
\mathcal{K}	Coeff. For reference histograms $h_{ref,1}^c, h_{ref,2}^c$ update	(0.1, 0.05)

Table 4. Characteristics and parameter values for our image-based upper human body parts tracker



Figure 9. Tracking scenario including two persons with target out-of-sight. Target loss detection and automatic re-initialization

Thank You for previewing this eBook

You can read the full version of this eBook in different formats:

- HTML (Free /Available to everyone)
- PDF / TXT (Available to V.I.P. members. Free Standard members can access up to 5 PDF/TXT eBooks per month each month)
- Epub & Mobipocket (Exclusive to V.I.P. members)

To download this full book, simply select the format you desire below

