# 19

# Towards an Interactive Humanoid Companion with Visual Tracking Modalities

Paulo Menezes[1], Frédéric Lerasle[2, 3], Jorge Dias[1] and Thierry Germa[2]
[1]*Institute of Systems and Robotics - University of Coimbra,* [2]*LAAS-CNRS,* [3]*Université Paul Sabatier*
[1]*Portugal,* [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



Source: Humanoid Robots: Human-like Machines, Book edited by: Matthias Hackel
ISBN 978-3-902613-07-3, pp. 642, Itech, Vienna, Austria, June 2007

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 objets (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; Urtasum 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. 2993; 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; Urtasum 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 $z_{1:k} = z_1, \ldots, z_k$ through the linear point-mass combination

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

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

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 $q(.)$, which aims at adaptively exploring ``relevant'' areas of the state space. Their weights $\omega_k^{(i)}$ 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 $x_0^{(1)}, \ldots, x_0^{(i)}, \ldots, x_0^{(N)}$ 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** $i = 1, \ldots, N$, **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 $\{1, \ldots, N\}$ the indexes $s^{(1)}, \ldots, s^{(N)}$ 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 $q(x_k|x_{k-1}^{(i)}, z_k) = p(x_k^{(i)}|x_{k-1}^{(i)})$. 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 $q(.)$, 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 the measurements, dynamics and the prior, so that, with $\alpha, \beta \in [0; 1]$

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

### 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.

$[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^N$ = AUXILIARY_PF($[\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^N, z_k$)

1: **IF** $k = 0$, **THEN** Draw $x_0^{(1)}, \ldots, x_0^{(i)}, \ldots, x_0^{(N)}$ 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** $i = 1, \ldots, N$, **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 $\mu_k^{(i)} = \mathrm{E}_{p(x_k|x_{k-1}^{(i)})}[x_k]$ –, 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 $\{1, \ldots, N\}$ the indexes $s^{(1)}, \ldots, s^{(N)}$ 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 $[\{x_{k-1}^{(s^{(i)})}, \frac{1}{N}\}]_{i=1}^N$ ; both $\sum_{i=1}^N \lambda_k^{(i)} \delta(x_{k-1} - x_{k-1}^{(i)})$ and $\frac{1}{N}\sum_{i=1}^N \delta(x_{k-1} - x_{k-1}^{(s^{(i)})})$ approximate the smoothing pdf $p(x_{k-1}|z_{1:k})$
7: **FOR** $i = 1, \ldots, N$, **DO**
8: "Propagate" the particles by independently drawing $x_k^{(i)} \sim p(x_k|x_{k-1}^{(s^{(i)})})$
9: Update the weights, prior to their normalization, by setting $w_k^{(i)} \propto \frac{p(z_k|x_k^{(i)})}{\hat{p}(z_k|x_{k-1}^{(s^{(i)})})} = \frac{p(z_k|x_k^{(i)})}{p(z_k|\mu_k^{(s^{(i)})})}$
10: Compute the conditional mean of any function of $x_k$, *e.g.* the MMSE estimate $\mathrm{E}_{p(x_k|z_{1:k})}[x_k]$, 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 $\pi(.)$ 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.

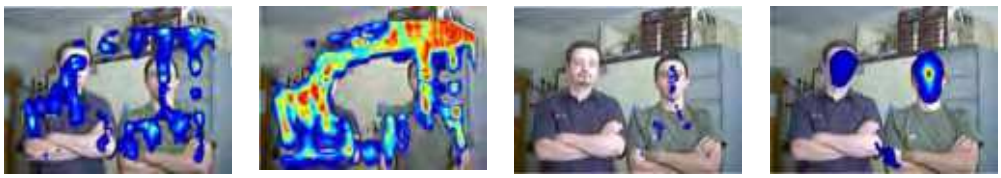
$$p(z^S|x) \propto \exp\left(\lambda_s \frac{D^2}{2\sigma_s^2}\right), \; D = \frac{1}{N_p}\sum_{j=1}^{N_p}|x(j) - z(j)|. \tag{2}$$

where $\lambda_s$ is a weight dedicated to further 3D tracking purpose (see section 6.2), j indexes the $N_p$ model points, and $\sigma_s$ 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

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

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 $N \geq 100$.



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

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 $\vec{I}(z(j))$ must have nonzero norm. The new likelihood $p(z^{MS}|x)$ involves the following similarity measure

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

where $\gamma(z(j)) = 0$ (resp. 1) if $\vec{I}(z(j)) \neq 0$ (resp. if $\vec{I}(z(j)) = 0$) 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

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

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 $h^c_{ref}$ and $h^c_x$ 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 $h^c_x$ close to the reference histogram $h^c_{ref}$. 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 $\{h^c_{p,ref} : c \in \{R,G,B\}, p = 1,\ldots,N_R\}$. By assuming conditional independence of the colour measurements, the likelihood $p(z^C|x)$ becomes

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

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 $(z^1,\ldots,z^M)$, the global measurement function can be factorized as

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

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.

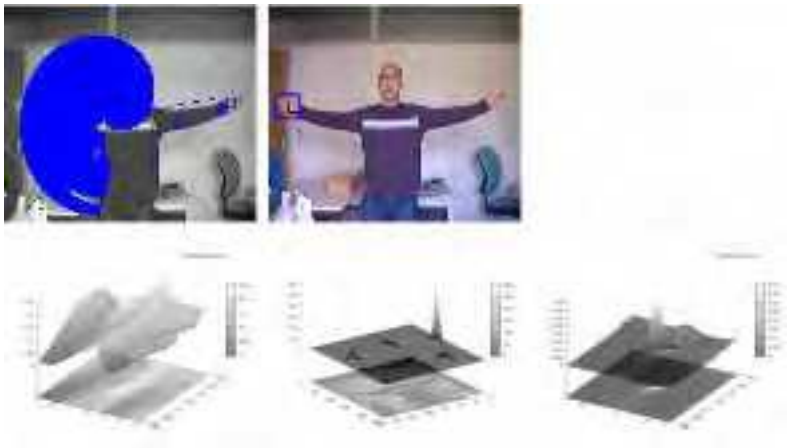


Figure 3. Likelihood plots for 3D tracking: shape cue, colour cue, both shape and colour cues

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 $\pi(.)$ 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).

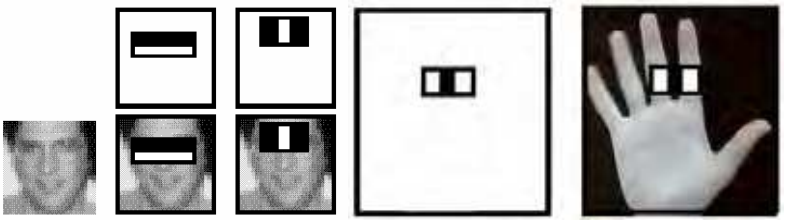


Figure 4. First Haar features for faces (Viola et al., 2001) and for hands

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 $\mathbf{p}_i = (u_i, v_i), i = 1, \ldots, N_B$ 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

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

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

#### 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 $\pi(\mathbf{x}|z^C)$ 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 $\pi(.)$ can be extended to consider the outputs from any of the M detectors, i.e.

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

## 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 $\theta$. 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 $\mathbf{x}_k = (u_k, v_k, s_k, \theta_k)^T$ 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 $\mathcal{N}(.|\mu, \Sigma)$ terms the Gaussian distribution with mean $\mu$ and covariance $\Sigma = diag(\sigma_u^2, \sigma_v^2, \sigma_s^2, \sigma_\theta^2)$

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 $p(z^{MS}|x)$ 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) |
| $\sigma_s$ | standard deviation in $p(z^{MS}/x_k)$ | 36 |
| $N_p$ | Number of model points in similarity measure | 30 |
| $\rho$ | 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 $C_l$ out of the set $\{C_l\}_{1 \leq l \leq M}$ -- 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.

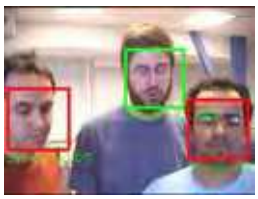


Figure 6. Example of face recognition

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 $\{\mathcal{F}(j), j \in \{1,...30\times30\}\}$, the DFFS criteria is written as follows

$$\mathscr{D} = \sum_{j=1}^{30\times30} |\mathcal{F}(j) - \mathcal{F}_r(j)| \quad (10)$$

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 $\mathscr{D}_l = \mathscr{D}(\mathcal{F}, C_l)$ and an a priori probability $P(C_l|\mathcal{F})$ of labelling to $C_l$

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

where $\mathcal{T}$ is a threshold predefined automatically, $C_\emptyset$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 $\mathscr{D}$ . 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).

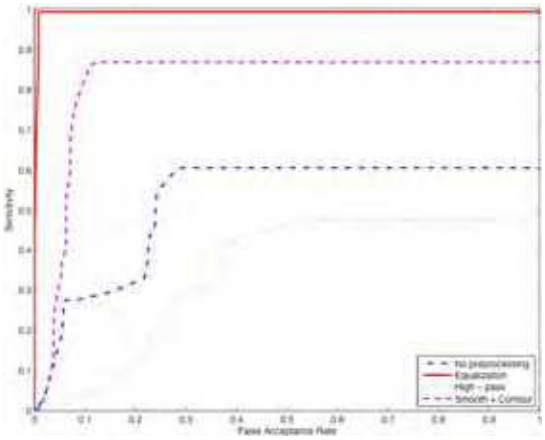


Figure 7. ROC curves image for different pre-processing techniques

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 $\mathbf{x}_k = [u_k, v_k, s_k]^T$ - the orientation $\theta_k$ 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

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

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 $P(C_l|\mathcal{F})$ probabilities egal 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 $P(C_l|\mathcal{F}_j) = 1$. 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.

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

where $\kappa$ 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 ($> |45°|$), 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) |
| $(\alpha, \beta)$ | Coeff. in importance function $q(\mathbf{x}_k \mid \mathbf{x}_{k-1}, z_k)$ | (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 $\pi(\mathbf{x}_k \mid z^S)$ | (6, 6) |
| $(\sigma_{u_i}, \sigma_{v_i})$ | Standard deviation in importance function $\pi(\mathbf{x}_k \mid z^C)$ | (6, 6) |
| $N_p$ | Number of model points in similarity measure | 15 |
| $\sigma_s$ | Standard deviation in shape-based likelihood $p(z_k^S \mid \mathbf{x}_k)$ | 1.5 |
| $N_R$ | Number of patches in $p(z_k^C \mid \mathbf{x}_k)$ | 2 |
| $\sigma_C$ | Standard deviation in color based likelihood $p(z_k^C \mid \mathbf{x}_k)$ | 0.03 |
| $N_{bi}$ | Number of colour bins per channel involved in $p(z_k^C \mid \mathbf{x}_k)$ | 32 |
| $\kappa$ | Coeff. For reference histograms $h^c_{ref,1}, h^c_{ref,2}$ 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

The first scenario (Figure 9), involving sporadic target disappearance, shows that our probabilistic tracker is correctly positioned on the desired person (on the background) throughout the sequence. Although the later disappears, the tracker doesn't lock onto the undesired person thanks to low recognition likelihood. The tracker re-initializes automatically as soon as the target re-appears.
The second scenario (Figure 10) involves occlusion of the target by another person traversing the field of view. The combination of multiple cues based likelihood and face recognition allows keeping track of the region of interest even after the complete occlusion. These videos as well as other sequences are available at the URL http://www.laas.fr/tgerma/videos demonstrate the tracker's performances on a number of challenging scenarios. This tracker relates to the module ICU in the Jido's software architecture (see section 7.1).

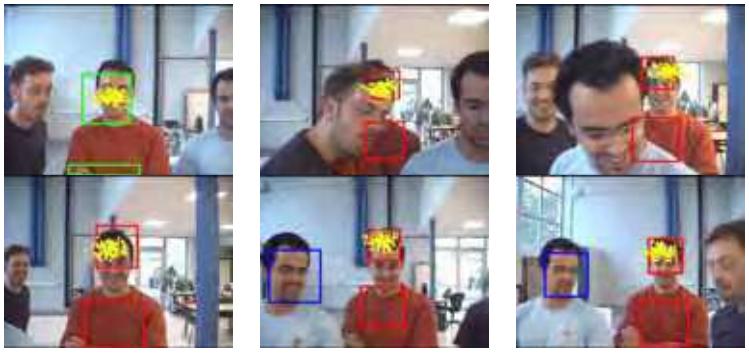


Figure 10. Tracking scenario involving full occlusions between persons

## 6. 3D tracking dedicated to upper human body parts

### 6.1 Preliminary works for hands tracking

#### 6.1.1 Basis on quadrics projection

Manipulation tasks, e.g. exchanging object, rely on 3D tracking of the human hands. Therefore, preliminary developments deals with an appearance and 3D model based tracker of hands from a mobile platform. This involves the perspective projection of the associated 3D model. Hands are modelled by deformable quadrics as quadrics and truncated ones can represent approximatively most of the upper human body parts (arm, forearm, torso,...). Let us recall some basis relative to the projection of quadrics. The projection matrix of a quadric in a normalised camera is $\mathbf{P} = [\mathbb{I}_{3\times2}|\mathbf{0}_{3\times1}]$. Considering a pinhole camera model, the camera centre and an image point $\mathbf{X}$ define a projective ray, which contains the points given by $\mathbf{X} = [\mathbf{x}\ s]^T$. The depth of a 3D point, situated on this ray, is determined by the scalar. The expression of the quadric $\mathbf{X}^T\mathbf{Q}\mathbf{X} = 0$ can be written as

$$\mathbf{x}^T\mathbf{A}\mathbf{x} + 2s\mathbf{b}^T\mathbf{x} + s^2c = 0, \quad \text{where} \quad \mathbf{Q} = \begin{bmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^T & c \end{bmatrix}. \tag{14}$$

This expression can be considered as an equation of second degree in s. Then, when the ray represented by $\mathbf{X}(s)$ is tangent to $\mathcal{Q}$, there is a single solution for equation (14) i.e. its discriminant is zero, so:

$$\mathbf{x}^T(\mathbf{b}\mathbf{b}^T - c\mathbf{A})\mathbf{x} = 0. \quad (15)$$

This expression corresponds to a conic $\mathcal{C}$ in the image plane, with $\mathbf{C} = c\mathbf{A} - \mathbf{b}\mathbf{b}^T$, which therefore corresponds to the projection of the quadric $\mathcal{Q}$. For any $\mathbf{x}$ belonging to $\mathcal{C}$, the corresponding 3D point $\mathbf{X}$ is fully defined by finding $s_0$ which is given by $s_0 = -\mathbf{b}^T\mathbf{x}/c$. This formula can be extended to arbitrary projective camera with $\mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}]$. Defining a 4x4 matrix $\mathbf{H}$ such that $\mathbf{PH} = [\mathbb{I}|\mathbf{0}]$, and then $\mathbf{x} = [\mathbb{I}|\mathbf{0}]\mathbf{H}^{-1}\mathbf{X}$. Only three types of quadrics are of interest for the next: ellipsoid, cone and cylinder. Hands are here modelled by ellipsoids which image projection for a pinhole camera is an ellipse.

### 6.1.2 Implementation and results

Our quadric model is deformable to deal with the shape appearance variations depending on the configuration in space. For a targeted hand, the state vector entries relate to the position $(t_x, t_y, t_z)$, orientation $(\alpha, \beta, \gamma)$, and axis lengths $(\sigma_x, \sigma_y, \sigma_z)$. As was the case for the above trackers, the state vector entries are assumed to evolve according to mutually independent Gaussian random walk models. From these considerations, the state vector to estimate at time k is

$$\mathbf{x}_k = (t_{x_k}, t_{y_k}, t_{z_k}, \alpha_k, \beta_k, \gamma_k, \sigma_{x_k}, \sigma_{y_k}, \sigma_{z_k})'. \quad (16)$$

We adopt ICONDENSATION as this strategy permits to sample some particles on the target after sporadic disappearances from the field of view. This situation is frequently encountered during manipulation tasks as short human-robot distances ([0.5;1.5] m) are involved. Using the video stream issued from the binocular system embedded on the robot here enlarges the environment perception at such close-range distances. Let us characterize the importance and measurement functions.

A pre-processing step consists in determining correspondences based on heuristics (Schmidt 1996) for skin colour blobs (see 4.2) in the image pair and calculate the 3D position for each match explicitly by triangulation. The face blob is beforehand filtered thanks to the face detector. Given M extracted image blobs at time k, the j-th blob is represented by its centroid $m^L_{j,k}$ (resp. $m^R_{j,k}$) and inertia matrix $\sigma^L_{j,k}$ (resp. $\sigma^R_{j,k}$) in the left and right image. For each triangulated position $P_{j,k} = g(m^L_{j,k}, m^R_{j,k})$, the covariance matrix is estimated as follows

$$\Sigma_{j,k} = G^L_{j,k}.\sigma^L_{j,k}.{G^L_{j,k}}^T + G^R_{j,k}\sigma^R_{j,k}{G^R_{j,k}}^T, \quad (17)$$

where $G^L_{j,k}$ and $G^R_{j,k}$ are the appropriate Jacobian matrices relatively to g(.). Finally, SVD decomposition of matrix $\Sigma_{j,k}$ allows defining a new importance function $\pi(\mathbf{x}_k|z^{3d}_k)$, which related to 3D constraint. Thus, the hands can be understood as three natural markers aiming at positioning the particles during the importance sampling. Note that most blob correspondences between the image planes are correct during this pre-processing step, but that there are also a few outliers (typically the non-targeted hand) that are filtered afterwards in the tracking loop.

Regarding the measurement function, the 3D ellipsoids (corresponding to hypothesis after importance sampling) are projected into a single image plane thanks to (15). The global measurement function fuses shape and colour cues, and can then be formulated as: $p(z^S_k, z^C_k|\mathbf{x}_k) = p(z^S|\mathbf{x}_k).p(z^C_k|\mathbf{x}_k)$. The shape-based likelihood involves the similarity measure (2)

while the colour-based likelihood involves histogram updating thanks to (13). Characteristics and parameter values describing the likelihoods state dynamics are listed in Table 5.

| Symbol | Meaning | Value |
|---|---|---|
| - | Particle filtering strategy | ICONDENSATION |
| N | Number of particles | 100 |
| (nbL, nbC) | Image resolution | (320, 240) |
| $(\alpha, \beta)$ | Coeff. In importance function $q(\mathbf{x}_k \mid \mathbf{x}_{k-1}, z_k)$ | (0.3, 0.6) |
| $(\sigma_{t_x}, \sigma_{t_y}, \sigma_{t_z})$ | Standard deviation in random walk models | (1, 1, 1) cm |
| $(\sigma_\alpha, \sigma_\beta, \sigma_\gamma)$ | Standard deviation in random walk models | (1, 1, 1) degrees |
| $(\sigma_\alpha, \sigma_\beta, \sigma_\gamma)$ | Standard deviation in random walk models | $(10^{-4}, 10^{-4}, 10^{-4})$ |
| $N_p$ | Number of model points in similarity measure | 20 |
| $\sigma_s$ | Standard deviation in shape-based likelihood $p(z_k^S \mid \mathbf{x}_k)$ | 1.5 |
| $N_R$ | Number of patches in $p(z_k^C \mid \mathbf{x}_k)$ | 1 |
| $\sigma_c$ | Standard deviation in colour-based likelihood | 0.03 |
| $N_{bi}$ | Number of colour bins per channel involved in $p(z_k^C \mid \mathbf{x}_k)$ | 32 |
| $\kappa$ | Coeff. For reference histograms $h_{ref,1}^c$ update | 0.1 |

Table 5. Characteristics and parameter values for our 3D tracker of human hand

The particle filter is run with a set of N=100 particles. The computational time, processed on a 3GHz Pentium IV, is approximatively 25fps. Although tracking a single hand is meaningful when exchanging objects, our approach can be easily extended to track both the two hands.

Figure 11. 3D tracking of a hand in heavy clutter. The blue (resp. red) ellipses depict the particles (resp. the MMSE estimate). Automatic re-initialization after target loss

Figure 11 shows the ellipsoid projections on snapshots of a video sequence involving heavy clutter. As one can see, the ellipsoids are deformed, but their positions are still correct. Moreover, the tracker initializes itself automatically and recovers the target after transient failure. This 3D tracker relates to the module GEST in the Jido's software architecture (see section 7.1).

### 6.2 Extension to the upper human body limbs

This appearance-based tracker is devoted to tasks that rely on understanding/imitation of gestures or activities. We have extended the previous tracker to the whole upper human body parts. The upper human body parts are here modelled by rigid truncated conics. The method to handle model projection and hidden parts removal is inspired from section 6.1.1. Details can be found in (Menezes et al., 2005a). Our emphasis in the paper concerns the characterization of a new measurement model that combines edges and motion cues, with also local color and texture patches on clothing or relative to the hands acting as natural markers. One important point is that our approach could be easily scalable from one single to multiple views as well as to higher number of degrees of freedom (DOF), although it will introduce the consequent increase in computational cost. The following subsections detail some geometric properties the model entailed in the measurement model and the tracker implementation.

#### 6.2.1 Non-observable part stabilisation

Despite the visual cues depicted in section 4, ambiguities arise when certain model parameters cannot be inferred from the current image observations, especially for a monocular system. They include, but are not limited to, kinematical ambiguities. For instance, when one arm is straight and the shape-based likelihood (2) is used, rotation of the upper arm around its axial axis is unobservable, because the model's projected contours remain static under this DOF. Leaving these parameters unconstrained is questionable. Consequently, and like in (Sminchisescu et al, 2003), we control these parameters with a stabiliser cost function that reaches its minimum on a predefined rest configuration $x_{def}$. This enables the saving of computing efforts that would explore the unobservable regions of the configuration space. In the absence of strong observations, the parameters are constrained to lie near their default values, whereas strong observations unstick the parameters values from these default configurations. The likelihood function for a state x is defined as $p_{st}(x) \propto \exp(-\lambda_{st}||x_{def} - x||^2)$. This prior depends on the structure parameters and the factor $\lambda_{st}$ will be chosen so that the stabilising effect will be negligible for the whole configuration space with the exception of the regions where the other cost terms are constant.

#### 6.2.2 Collision detection

Physical consistency imposes that the different body parts do not interpenetrate. As the estimation is based on a search on the configuration space it would be desirable to a priori remove those regions that correspond to collisions between parts. Unfortunately it is in general not possible to define these forbidden regions in closed form so they could be rejected immediately during the sample phase. The result is that in the particle filter framework, it is possible that configurations proposed by some particles correspond to such impossible configurations, thus exploring regions in the configuration space that are of no interest. To avoid these situations, we use a binary cost function that is not related to observations but only based on a collision detection mechanism. The likelihood function for a state x is $p_{coll}(x) \propto \exp(-\lambda_{co} f_{co})$ with:

$$f_{co}(x) = \begin{cases} 0 & \text{No collision} \\ 1 & \text{In collision} \end{cases}. \tag{18}$$

This function, although being discontinuous for some points of the configuration space and constant for all the remaining, is still usable in a Dirac particle filter context. The advantage

of its use is twofold. First it avoids the derivation of the filter to zones of no interest, and second it avoids wasting time in performing the measuring step for unacceptable hypothesis as they can be immediately rejected.

### 6.3 Observation likelihoods and model priors fusion

Fusing multiple visual cues enables the tracker to better benefit from M distinct measurements $(z^1, \ldots, z^M)$. Assuming that these are mutually independent conditioned on the state, and given L model priors $(p_1(x), \ldots, p_L(x)$, the unified measurement function thus factorizes as

$$p(z^1, \ldots, z^M|x) \propto \prod_{i=1}^{M} p(z^i|x). \prod_{j=1}^{L} p_j(x) \tag{19}$$

#### 6.3.1 Implementation and experiments

In its actual form, our 3D tracker deals two-arms gestures under an 8-DOF model, i.e. four per arm. We assume therefore that the torso is coarsely fronto-parallel with respect to the camera while the position of the shoulders are deduced from the position of the face given by dedicated tracker (Menezes et al., 2004b). All the DOFs are accounted for in the state vector $\mathbf{X}k$ related to the k-th frame. Kinematical constraints require that the values of these joint angles evolve within anatomically consistent intervals. Samples (issued from the proposal) falling outside the admissible joint parameter range are enforced to the hard limits and not rejected as this could lead to cascades of rejected moves that slow down the sampler.

Recall that pdf $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ encodes information about the dynamics of the targeted human limbs. These are described by an Auto-Regressive model with the following form $\mathbf{x}_k^{'} = A\mathbf{x}_{k-1}^{'} + \mathbf{w}_k$ where $\mathbf{x}_k^{'} = [\mathbf{x}_k, \mathbf{x}_{k-1}]^{'}$, and $\mathbf{w}_k$ defines the process noise. In the current implementation, these dynamics correspond to a constant velocity model. We find this AR model gives empirically better results than usual random walk model, although a more rigorous evaluation would be here desirable. The patches are distributed on the surface model and their possible occlusions are managed during the tracking process. Our approach is different from the traditional marker-based ones because we do not use artificial but natural colour or texture-based markers e.g. the two hands and ROIs on the clothes. We adopt the APF scheme (Table 2), which allows to use some low cost measure or a priori knowledge to guide the particle placement, therefore concentrating them on the regions of interest of the state space. The measurement strategy is as follows: (1) particles are firstly located in good places of the configuration space according to rough correspondences between virtual patches related to the two hands projection and skin-like image ROIs involving likelihood $p(z_k^C|\mathbf{x}_k)$ in step 6, (2) particles' weights are fine-tuned by combining shape and motion cues ($p(z_k^{MS}|\mathbf{x}_k)$ entailing similarity distance (5), three color patches per arm (likelihood (6)) as well as model priors thanks to (19) in step 9. A second and important line of investigation concerns the incorporation of appropriate degrees of adaptability into these multiple cues based likelihoods depending on the target appearance and environmental conditions. Therefore, some heuristics allows weighting the strength of each visual cue in the unified likelihood (7). An a priori confidence criterion of a given coloured or textured patch relative to clothes can be easily derived from the associated likelihood functions where the p-th colour reference histogram $h_{ref,p}^c$ ($h_{ref,p}^t$ for texture one) is uniform

and so given by $h^c_{j,ref} = \frac{1}{N_{bi}},\ j = 1,\ldots,N_{bi}$ where index p has been omitted for compactness reasons. Typically, uniform coloured patches produce low likelihood values, whereas higher likelihood values characterise confident patches because their associated colour distributions are discriminant and ensure non-ambiguous matching. As stated before, parameters $\lambda_{p,c}$ and $\lambda_{p,t}$ weight the strength of the p-th marker in the likelihood (19). In the same way, the parameter $\lambda_s$weights the edges density contribution and is fixed from the first DT image of the sequence.

| Symbol | Meaning | Value |
|---|---|---|
| - | Particle filtering strategy | APF |
| N | Number of particles | 400 |
| (nbL, nbC) | Image resolution | (320, 240) |
| $\lambda_{st}$ | Factor in model prior $p_{st}(x)$ | 0.5 |
| $\lambda_{co}$ | Factor in model prior $p_{co}(x)$ | 0.5 |
| K | Penalty in similarity distance | 0.5 |
| $\sigma_s$ | Standard deviation in similarity distance | 1 |
| $N_R$ | Number of patches in color-based likelihood $p(z_k^C \mid \mathbf{x}_k)$ | 6 |
| $\sigma_C$ | Standard deviation in$p(z_k^C \mid \mathbf{x}_k)$ | 0.3 |
| $N_{bi}$ | Number of color bins per channel involved in $p(z_k^C \mid \mathbf{x}_k)$ | 32 |

Table 6. Characteristics and parameter values of our upper human body 3D tracker

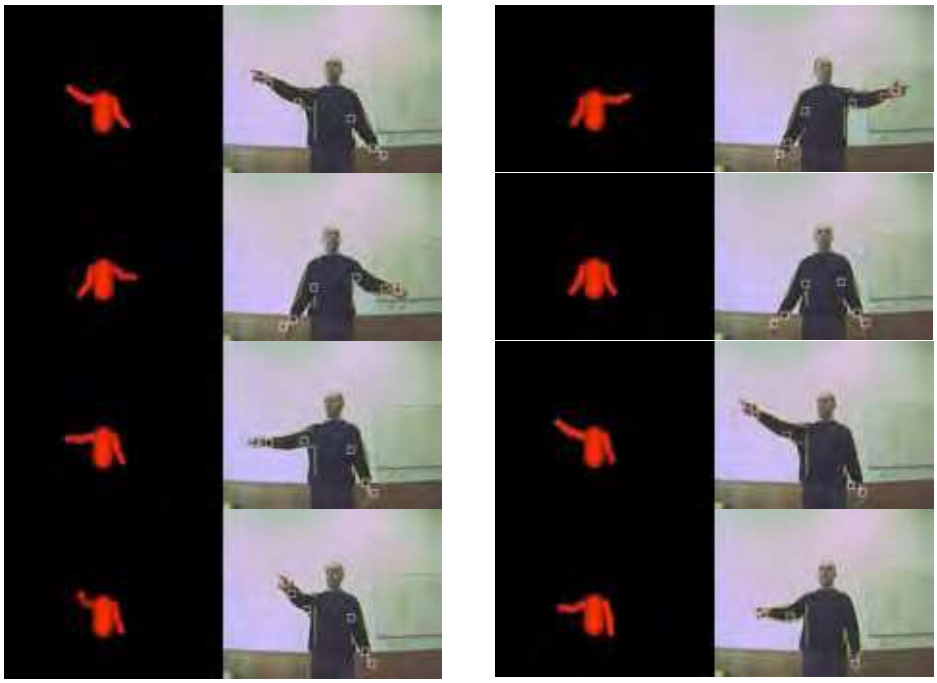


Figure 12. From top-left to bottom right: snapshots of tracking sequence involving deictic gestures

Due to the efficiency of the importance density and the relatively low dimensionality of the state-space, tracking results are achieved with a reasonably small number of particles i.e. N=400 particles. In our non-optimised implementation, a PentiumIV-3GHz requires about 1fps to process the two arm tracking, most of the time being spent in evaluating the observation function. To compare, classic systems take a few seconds per frame to process a single arm tracking. Characteristics and parameter values involved in the likelihoods, proposal and state dynamics of our upper human body 3D tracker are reported in Table 6.
The above-described approach has been implemented and evaluated over monocular images sequences. To illustrate our approach, we show and comment snapshots from two typical sequences acquired from the Jido robot (see section 7) in different situations to highlight our adaptative data fusion strategy. The full images as well as other sequences can be found at the URL http://www.isr.uc.pt/~paulo/HRI . The first sequence (Figure 12) was shot against a white and unevenly illuminated background, but involves loose fitting closing and near fronto-parallel motions. The second sequence (Figure 13) involves coarse fronto-parallel motions over a heavy cluttered background. For each snapshot, the right sub-figures overlay the model projections on to the original images for the MMSE estimate, while the left ones show its corresponding estimated configuration corresponding to the posterior pdf $p(\mathbf{x}_k|z_{1:k})$.

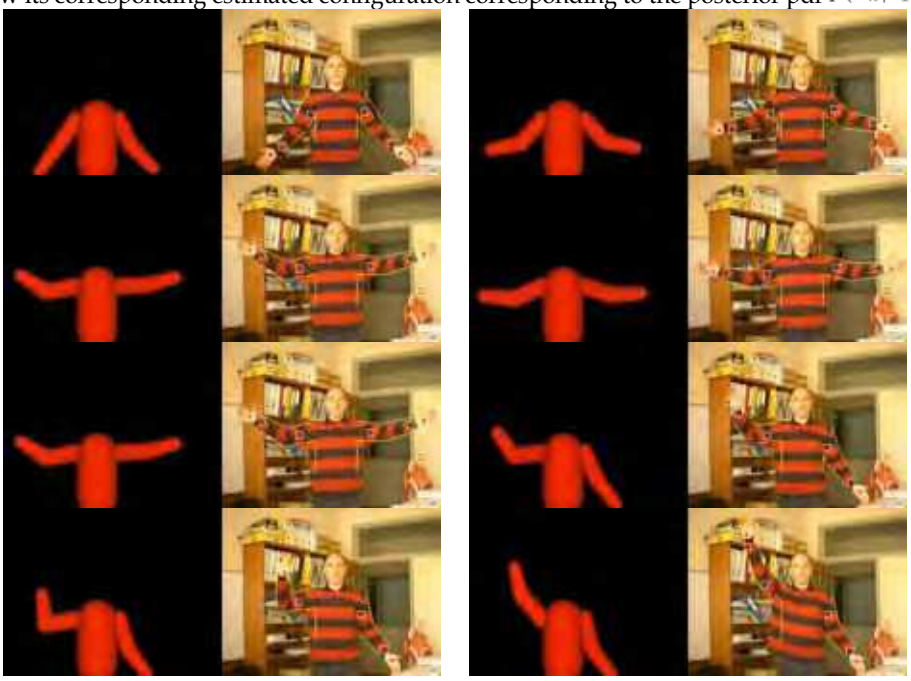


Figure 13. From top-left to bottom-right: snapshots of tracking sequence involving heavy clutter

The first tracking scenario (Figure 12) involves pointing gestures. The target contours are prominent and are weakly disturbed by the background clutter. The high confident contours cue ensures the tracking success. The patches on the uniform sweet are here of little help, as their enclosed colour or texture distributions are quite uniform. The adaptative

system allows giving them a weak strength in the unified likelihood cost ($\lambda_{p,c} = 0.1$ against $\lambda_s = 10$). They do not introduce any improvement with respect to their position on the arm, but their benefit comes in the form of inside/outside information, which complements the contours especially when they failed. This permitted the tracking of the arms even when they got out of the fronto-parallel plane thanks to all the patches (Figure 12).
For the second scenario (Figure 13), the tracker deals with significantly more complex scene but tracks also the full sequence without failure. This scenario takes clearly benefit from the introduction of discriminant patches as their colour distributions are far from uniform ones. This leads to higher values of confidence dedicated to the likelihood $p(z_k^c / x_k)$, namely $\lambda_{p,c} = 1$. In these challenging operating conditions, two heuristics allow jointly to release from distracting clutter that might partly resemble human body parts (for instance the cupboard pillar[1]). On the one hand, estimating the edges density in the first frame highlights that shape cue is not a confident one in this context, so its confidence level in the global cost (19) is reduced accordingly during the tracking process i.e. $\lambda_s = 0.1$. On the other hand, optical flow weights the importance relative to the foreground and background contours thanks to the likelihood $p(z_k^{MS}|\mathbf{x}_k)$. If considering only contour cues in the likelihood, the tracker would attach itself to cluttered zones and consequently lose the target. This tracker relates to the module TPB in the Jido's software architecture (see section 7.1).

## 7. Integration on robotic platforms dedicated to human-robot interaction

### 7.1 Integration on a robot companion

#### 7.1.1 Outline of the overall software architecture

The above visual functions were embedded on a robot companion called Jido. Jido is equipped with: (i) a 6-DOF arm, (ii) a pan-tilt stereo system at the top of a mast (dedicated to human-robot interaction mechanisms), (iii) a second video system fixed on the arm wrist for object grasping, (iv) two laser scanners, (v) one panelPC with tactile screen for interaction purpose, (vi) one screen to provide feedback to the robot user. Jido has been endowed with functions enabling to act as robot companion and especially to exchange objects with human beings. So, it embeds robust and efficient basic navigation and object recognition abilities. Besides, our efforts focuses in this article concern the design of visual functions in order to recognize individuals and track his/her human body parts during object exchange tasks.
To this aim, Jido is fitted with the "LAAS" layered software architecture thoroughly presented in (Alami et al., 1998). On the top of the hardware (sensors and effectors), the functional level listed in Figure 14, encapsulates all the robot's action and perception capabilities into controllable communicating modules, operating at very strong temporal constraints. The executive level activates these modules, controls the embedded functions, and coordinates the services depending on the task high-level requirements. Finally, the upper decision level copes with task planning and supervision, while remaining reactive to events from the execution control level. The integration of our visual modalities (green boxes) is currently carried out in the architecture, which resides on the Jido robot.
The modules GEST, HumRec, and ICU have been fully integrated in the Jido's software architecture. The module TBP has been devoted preliminary to the HRP2 model (see section

[1] with also skin-like color...

7.2). Before integration in Jido, we aim beforehand at extending this tracker to cope with stereoscopic data (Fontmarty et al., 2007).

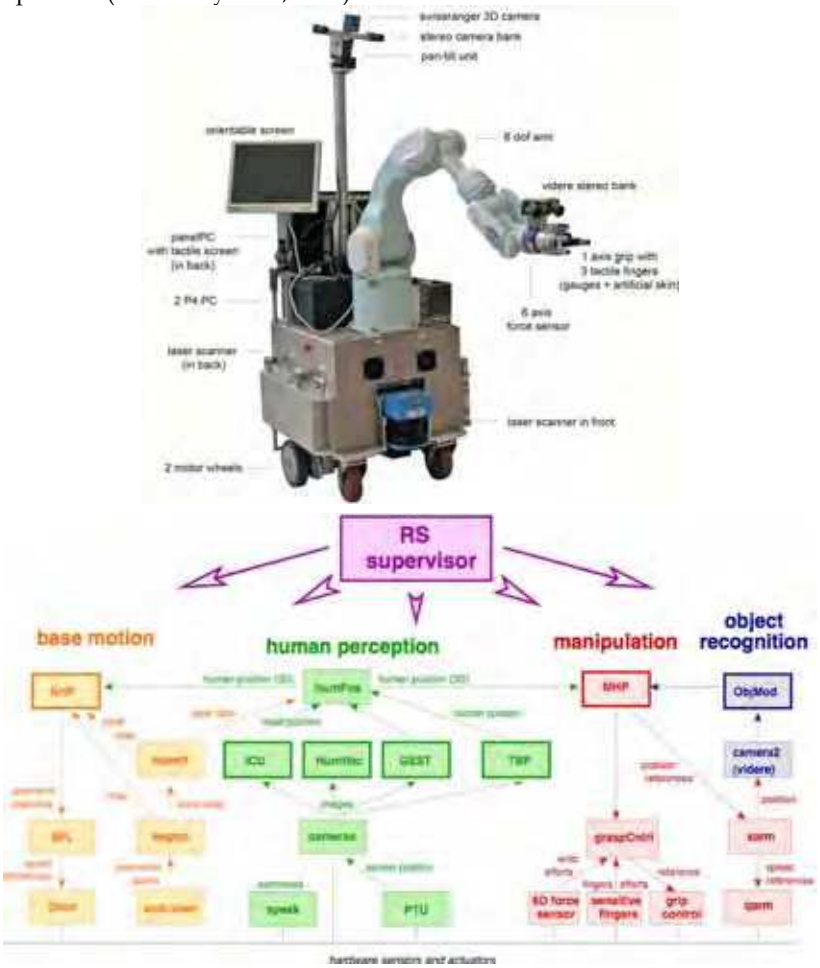



Figure 14. Jido robot and its layered software architecture

### 7.1.2 Considerations about the visual modalities software architecture

The C++ implementation of the modules are integrated in the ``LAAS'' architecture using a C/C++ interfacing scheme. They enjoy a high modularity thanks to C++ abstract classes and template implementations. This way, virtually any tracker can be implemented by selecting its components from predefined libraries related to particle filtering strategies, state evolution models, and measurement / importance functions. For more flexibility, specific components can be defined and integrated directly. A finite-state automaton can be designed from the vision-based services outlined in section 1. As illustrated in Figure 15, its states are respectively associated to the INIT mode and to the aforementioned vision-based modules while the arrows relate to the transitions between them. Another complementary

modalities (blue ellipses), not yet integrated into the robot architecture, have been also added. Heuristics relying on the current human-robot distance, face recognition status, and current executed task (red rectangles) allow to characterize the transitions in the graph. Note that the module ICU can be invoked from all the mentioned human-robot distances ([1;5]m.).

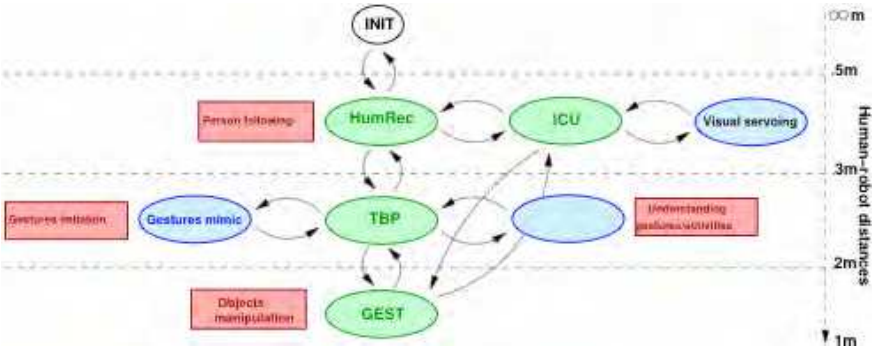



Figure 15. Transitions between vision-based modules

### 7.2 Integration on a HRP2 model dedicated to gestures imitation

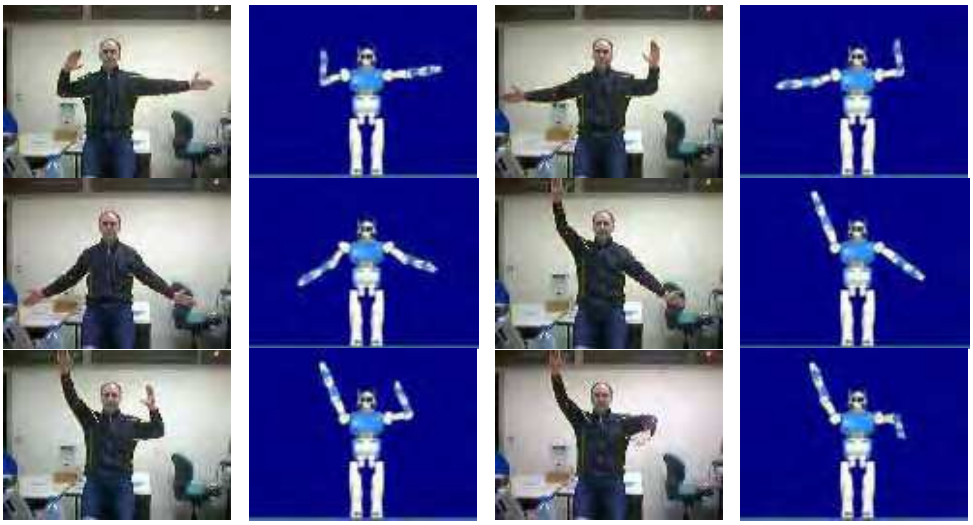


Figure 16. From top-left to bottom-right: snapshots of tracking sequence and animation of HRP2 using the estimated parameters

As mentioned before, a last envisaged application concerns gestures imitation by a humanoid robot (Menezes et al., 2005a). This involves 3D tracking of the upper human body limbs and mapping the joints of our 3D kinematical 3D model to those of the robot. In addition to the previous commented sequences, this scenario (Figure 16) with moderate clutter explores 3D estimation behaviour with respect to problematic motions i.e. non-fronto-parallel ones, elbow end-stops and observation ambiguities. The left column

represents the input images and the projection of the model contours superimposed while the right column represents the animation of the HRP2 using the estimated parameters[2]. The first frames involve both elbow end-stops and observation ambiguities. These particular configurations are easily dealt with in our particle-filtering framework. When elbow end-stop occurs, the sampler is able to maintain the elbow angle within its predefined hard limits. Observation ambiguity arises when the arm is straight. The twist parameter is temporary unobservable but remains stable thanks to the likelihood $p_{st}(\mathbf{x}_k)$. As highlighted in (Deutscher et al., 1999), Kalman filtering is quite unable to track through end-stop configurations. Some frames later in Figure 16, the left arm bends slightly towards the camera. Thanks to the patches on the hands, the tracker manages to follow this temporary unobservable motion, although it significantly misestimates the rotation during this motion. The entire video is available at http://www.isr.uc.pt/~paulo/HRI.

## 8. Conclusion

This article presents the developments of a set of visual trackers dedicated to the upper human body parts. We have outlined visual trackers a universal humanoid companion should deal with in the future. A brief state-of-art related to tracking highlight that particle filtering is widely used in the literature. The popularity of this framework stems, probably, from its simplicity, ease of implementation, and modelling flexibility, for a wide variety of applications.

From these considerations, a first contribution relates to visual data fusion and particle filtering strategies associations with respect to considered interaction modalities. This guiding principle frames all the designed and developed trackers. Practically, the multi-cues associations proved to be more robust than any of the cues individually. All the trackers are applied in quasi-real-time process and have the ability to (re)-initialize automatically.

A second contribution concerns especially the 3D tracker dedicated to the upper human body parts. An efficient method (not detailed here, see (Menezes et al., 2005b) has been proposed in order to handle the projection and hidden removal efficiently. In the vein of the depicted 2D trackers, we propose a new model-image matching cost metric combining visual cues but also geometric constraints. We integrate degrees of adaptability into this likelihood function depending on the human limbs appearance and the environmental conditions. Finally, integration, even if in progress, of the developed trackers on two platforms highlights their relevance and complementarity. To our knowledge, quite few mature robotic systems enjoy such advanced capabilities of human perception.

Several directions are studied regarding our trackers. Firstly, to achieve gestures/activities interpretation, Hidden Markov Models (Fox et al., 2006) and Dynamic Bayesian Network (Pavlovic et al., 1999) are currently under evaluation and preliminary results are actually available. Secondly, we currently study how to extend our monocular-based approaches to account for stereoscopic data as most humanoid robot embed such exteroceptive sensor. Finally, we will integrate all these visual trackers on our new humanoid companion HRP2. The tracking functionalities will be made much more active; zooming will be used to actively adapt the focal lenght with respect to the H/R distance and the current robot status.

---

[2] This animation was performed using the KineoWorks platform and the HRP2 model by courtesy of AIST (GeneralRobotix).

## 9. Acknowledgements

The work described in this paper has received partial financial support from Fundação para a Ciência e Tecnologia through a scholarship granted to the first author.
Parts of it were conducted within the EU Integrated Project COGNIRON (``The Cognitive Companion'') and funded by the European Commission Division FP6-IST Future and Emerging Technologies under Contract FP6-002020. We want also to thank Brice Burger for implementation and integration involvement regarding the hand 3D tracker.

## 10. References

Alami, R.; Chatila, R.; Fleury, S. & Ingrand, F. (1998). An architecture for autonomy. *Int. Journal of Robotic Research* (IJRR'98), 17(4):315–337.

Arulampalam, S; Maskell, S.; Gordon, N. & Clapp, T. (2002). A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *Trans. on Signal Processing*, 2(50):174–188.

Asfour, T.; Gyafas, F.; Azad, P. & Dillman, R. (2006). Imitation learning of dual-arm manipulation tasks in humanoid robot. *Int. Conf. on Humanoid Robots* (HUMANOID'06), pages 40–47, Genoa.

Barreto, J.; Menezes, P. & Dias, J. (2004). Human robot interaction based on haar-like features and eigenfaces. *Int. Conf. on Robotics and Automation* (ICRA'04), New Orleans.

Bennewitz, M.; Faber, F.; Joho, D.; Schreiber, M. & Behnke, S. (2005). Towards a humanoid museum guide robot that interacts with multiple persons. Dans *Int. Conf. on Humanoid Robots* (HUMANOID'05), pages 418–424, Tsukuba.

Brèthes, L.; Lerasle, F. & Danès, P. (2005). Data fusion for visual tracking dedicated to human-robot interaction. Int. Conf. on Robotics and Automation (ICRA'05). pages 2087-2092, Barcelona.

Chen, H. & Liu, T. (2001). Trust-region methods for real-time tracking. Dans *Int. Conf. on Computer Vision* (ICCV'01), volume 2, pages 717–722, Vancouver.

Comaniciu, D.; Ramesh, V. & Meer, P. (2003). Kernel-based object tracking. *Trans. on Pattern Analysis Machine Intelligence* (PAMI'03), volume 25, pages 564–575.

Delamarre, Q.; & Faugeras, O. (2001). 3D articulated models and multi-view tracking with physical forces. *Computer Vision and Image Understanding* (CVIU'01), 81:328–357.

Deutscher, J.; Blake, A. & Reid, I. (2000). Articulated body motion capture by annealed particle filtering. *Int. Conf. on Computer Vision and Pattern Recognition* (CVPR'00), pages 126–133, Hilton Head Island.

Deutscher, J.; Davison, A. & Reid, I. (2001). Automatic partitioning of high dimensional search spaces associated with articulated body motion capture. Dans *Int. Conf. on Computer Vision and Pattern Recognition* (CVPR'01), pages 669–676, Kauai.

Deutscher, J.; North, B.; Bascle, B. & Blake, A. (1999). Tracking trough singularities and discontinuities by random sampling. Dans *Int. Conf. on Computer Vision* (ICCV'99).

Doucet, A.; Godsill, S.J. & Andrieu, C. (2000). On sequential monte carlo sampling methods for bayesian filtering. S*tatistics and Computing*, 10(3):197–208.

Engelberger, J.; (1989). *Robotics in Service*, chap 1. Cambridge, MA; MIT Press.

Fitzpatrick, P.; Metta, G.; Natale, L.; Rao, S. & Sandini, G. (2003). Learning about objects through action-initial steps towards artificial cognition. *Int. Conf. on Robotics and Automation* (ICRA'03), pages 3140– 3145, Taipei, Taiwan.

Fong, T.; Nourbakhsh, I. & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems* (RAS'03), 42:143–166.

Fontmarty, M.; Lerasle, F. ; Danès, P. & Menezes, P. (2007). Filtrage particulaire pour la capture de mouvement dédiée à l'interaction homme-robot. *Congrès francophone ORASIS*, Obernai, France.

Fox, M.; Ghallab, M. ; Infantes, G. & Long, D. (2006). Robust introspection through learned hidden markov models. *Artificial Intelligence* (AI'06), 170(2):59–113.

Gavrila, D. (1996). 3D model-based tracking of human in actions: A multi-view approach. *Int. Conf. on Computer Vision and Pattern Recognition* (CVPR'96), pages 73-80, San Francisco.

Gavrila, D. (1999). The visual analysis of human movement: a survey. *Computer Vision and Image Understanding* (CVIU'99), 73(1):82–98.

Germa, T.; Brèthes, L.; Lerasle, F. & Simon, T. (2007). Data fusion and eigenface based tracking dedicated to a tour-guide robot. Dans *Association Internationale pour l'Automatisation Industrielle* (AIAI'07), Montréal, Canada. Int. Conf. On Vision Systems (ICVS'07), Bielefeld.

Giebel, J.; Gavrila, D. M. & Schnorr, C. (2004). A bayesian framework for multi-cue 3D object. *European Conf. on Computer Vision* (ECCV'04), Prague.

Goncalves, L.; Bernardo, E. D.; Ursella, E. & Perona, P. (1995). Monocular tracking of the human arm in 3D. *Int. Conf. on Computer Vision* (ICCV'95).

Heap, A. J. & Hogg, D. C. (1996). Towards 3D hand tracking using a deformable model. *Int. Conf. on Face and Gesture Recognition* (FGR'96), pages 140–145, Killington, USA.

Heseltine, T.; Pears, N. & Austin, J. (2002). Evaluation of image pre-processing techniques for eigenface based recognition. *Int. Conf. on Image and Graphics,* SPIE, pages 677–685.

Isard, M. & Blake, A. (1998a). CONDENSATION – conditional density propagation for visual tracking. *Int. Journal on Computer Vision* (IJCV'98), 29(1):5–28.

Isard, M. & Blake, A. (1998b). I-CONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. Dans *European Conf. on Computer Vision* (ECCV'98), pages 893–908.

Isard, M. & Blake, A. (1998c). A mixed-state condensation tracker with automatic model-switching. *International Conference on Computer Vision*, page 107, Washington, DC, USA. IEEE Computer Society.

Jones, M. & Rehg, J. (1998). Color detection. *Rapport technique*, Compaq Cambridge Research Lab.

Kakadiaris, I. & Metaxas, D. (2000). Model-based estimation of 3D human motion. *Trans. on Pattern Analysis and Machine Intelligence* (PAMI'00), 22(12):1453–1459.

Kehl, R. & Gool, L. (2004). Real-time pointing gesture recognition for an immersive environment. *Int. Conf. on Face and Gesture Recognition* (FGR'04), pages 577-582, Seoul.

Lerasle, F.; Rives, G. & Dhome, M. (1999). Tracking of human limbs by multiocular vision. *Computer Vision and Image Understanding* (CVIU'99), 75(3):229–246.

Maas, J. ; Spexard, T. ; Fritsch, J. ; Wrede, B. & Sagerer, G. (2006). BIRON, what's the topic? a multimodal topic tracker for improved human-robot interaction. *Int. Symp. on Robot and Human Interactive Communication* (RO-MAN'06), Hatfield, UK.

MacCormick, J. & Blake, A. (2000a). A probabilistic exclusion principle for tracking multiple objects. *Int. Journal of Computer Vision*, 39(1):57–71.

MacCormick, J. & Isard, M. (2000b). Partitioned sampling, articulated objects, and interface quality hand tracking. *European Conf. on Computer Vision* (ECCV'00), pages 3–19, London. Springer Verlag.

Menezes, P.; Barreto, J. & Dias, J. (2004). Face tracking based on haar-like features and eigenfaces. *IFAC Symp. on Intelligent Autonomous Vehicles*, Lisbon.

Menezes, P.; Lerasle, F.; Dias, J. & Chatila, R (2005a). Appearance-based tracking of 3D articulated structures. *Int. Symp. on Robotics* (ISR'05), Tokyo.

Menezes, P.; Lerasle, F.; Dias, J. & Chatila, R. (2005b). Single camera motions capture system dedicated to gestures imitation. *Int. Conf. on Humanoid Robots* (HUMANOID'05), pages 430–435, Tsukuba.

Metaxas, D. ; Samaras, D. & Oliensis, J. (2003). Using multiple cues for hand tracking and model refinement. *Int. Conf. on Computer Vision and Pattern Recognition* (CVPR'03), pages 443-450, Madison.

Nakazawa, A.; Nakaoka, S.; Kudo, S. & Ikeuchi, K. (2002). Imitating human dance motion through motion structure analysis. *Int. Conf. on Robotics and Automation* (ICRA'02), Washington.

Nourbakhsh, I.; Kunz C. & Willeke, D. (2003). The Mobot museum robot installations: A five year experiment. *Int. Conf. On Intelligent Robots and Systems* (IROS'03), Las Vegas.

Park, J.; Park, S. & Aggarwal, J. (2003). Human motion tracking by combining view-based and model-based methods for monocular vision. Dans *Int. Conf. on Computational Science and its Applications* (ICCSA'03), pages 650–659.

Pavlovic, V.; Rehg, J. & Cham, T. (1999). A dynamic bayesian network approach to tracking using learned switching dynamic models. Dans *Int. Conf. on Computer Vision and Pattern Recognition* (CVPR'99), Ft. Collins.

Pavlovic, V. ; Sharma, R. & Huang, T. S. (1997). Visual interpretation of hand gestures for human-computer interaction : A review. *Trans. On Pattern Analysis and Machine Intelligence* (PAMI'97), 19(7): 677–695.

Pérez, P.; Vermaak, J. & Blake, A. (2004). Data fusion for visual tracking with particles. *Proc. IEEE*, 92(3):495–513.

Pérez, P. ; Vermaak, J. & Gangnet, M. (2002). Color-based probabilistic tracking. Dans *European Conf. on Computer Vision* (ECCV'02), pages 661–675, Berlin.

Pitt, M. & Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446).

Poon, E. & Fleet, D. (2002). Hybrid monte carlo filtering: Edge-based tracking. Dans *Workshop on Motion and Video Computing*, pages 151–158, Orlando, USA.

Rehg, J. & Kanade, T. (1995). Model-based tracking of self-occluding articulated objects. *Int. Conf. on Computer Vision* (ICCV'95), pages 612–617, Cambridge.

Rui, Y. & Chen, Y. (2001). Better proposal distributions: Object tracking using unscented particle filter. *Int. Conf. on Computer Vision and Pattern Recognition* (CVPR'01), pages 786–793, Hawai.

Schmid, C. (1996). Appariement d'images par invariants locaux de niveaux de gris. *Thèse de doctorat*, Institut National Polytechnique de Grenoble.

Schwerdt, K. & Crowley, J. L. (2000). Robust face tracking using color. *Int. Conf. on Face and Gesture Recognition* (FGR'00), pages 90–95, Grenoble, France.

Shon, A. ; Grochow, K. & Rao, P. (2005). Imitation learning from human motion capture using gaussian process. *Int. Conf. on Humanoid Robots* (HUMANOID'05), pages 129–134, Tsukuba.

Sidenbladh, H. ; Black, M. & Fleet, D. (2000). Stochastic tracking of 3D human figures using 2D image motion. Dans *European Conf. on Computer Vision* (ECCV'00), pages 702–718, Dublin.

Siegwart, R. ; Arras, O.; Bouabdallah, S.; Burnier, D. ; Froidevaux, G.; Greppin, X.; Jensen, B.; Lorotte, A.; Mayor, L.; Meisser, M.; Philippsen, R.; Piguet, R.; Ramel, G.; Terrien, G. & N. Tomatis (2003). Robox at expo 0.2: a large scale installation of personal robots. *Robotics and Autonomous Systems* (RAS'03), 42:203–222.

Sminchisescu, C. & Triggs, B. (2003). Estimating articulated human motion with covariance scaled sampling. *Int. Journal on Robotic Research* (IJRR'03), 6(22):371–393.

Stenger, B.; Mendonça, P. R. S. & Cipolla, R. (2001). Model-based hand tracking using an unscented Kalman filter. *British Machine Vision Conf.* (BMVC'01), volume 1, pages 63–72, Manchester.

Stenger, B.; Thayananthan, A. ; Torr, P. & Cipolla, R. (2003). Filtering using a tree-based estimator. *Int. Conf. on Computer Vision (*ICCV'03), pages 1063–1070, Nice.

Sturman,D. & Zeltzer, D. (1994). A survey of glove-based input. *Computer Graphics and Applications*, 14(1):30–39.

Thayananthan, A.; Stenger, B.; Torr, P. & R. Cipolla (2003). Learning a kinematic prior for tree-based filtering. Dans *British Machine Vision Conf.* (BMVC'03), volume 2, pages 589–598, Norwick.

Torma, P. & Szepesvari, C. (2003). Sequential importance sampling for visual tracking reconsidered. Dans *AI and Statistics*, pages 198–205.

Thrun, S.; Beetz, M.; Bennewitz, M.; Burgard, W.; Cremers, A.B.; Dellaert, F.; Fox, D.; Halnel, D.; Rosenberg, C.; Roy, N.; Schulte, J. & Schultz, D. (2000). Probabilistic algorithms and the interactive museum tour-guide robot MINERVA. *Int. Journal of Robotics Research* (IJRR'00).

Turk, M. & Pentland, A. (1991). Face recognition using eigenfaces. *Int. Conf. on Computer Vision and Pattern Recognition* (CVPR'91), pages 586–591, Hawai.

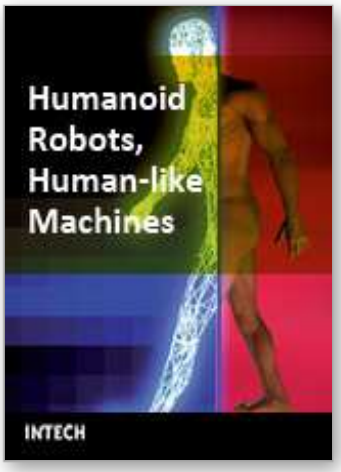
Urtasum, R. & Fua, P. (2004). 3D human body tracking using deterministic temporal motion models. *European Conf. on Computer Vision* (ECCV'04), Prague.

Viola, P. & Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. *Int. Conf. on Computer Vision and Pattern Recognition* (CVPR'01), Hawai.

Wachter, S. & Nagel, S. (1999). Tracking persons in monocular image sequences. *Computer Vision and Image Understanding* (CVIU'99), 74(3):174–192.

Wu, Y. & Huang, T. (1999). Vision-based gesture recognition: a review. *International Workshop on Gesture-Based Communication,* pages 103–105, Gif-sur-Yvette.

Wu, Y.; Lin, T. & Huang, T. (2001). Capturing natural hand articulation. I*nt. Conf. on Computer Vision* (ICCV'01), pages 426–432, Vancouver.




**Humanoid Robots, Human-like Machines**

Edited by Matthias Hackel

ISBN 978-3-902613-07-3

Hard cover, 642 pages

**Publisher** I-Tech Education and Publishing

**Published online** 01, June, 2007

**Published in print edition** June, 2007

In this book the variety of humanoid robotic research can be obtained. This book is divided in four parts: Hardware Development: Components and Systems, Biped Motion: Walking, Running and Self-orientation, Sensing the Environment: Acquisition, Data Processing and Control and Mind Organisation: Learning and Interaction. The first part of the book deals with remarkable hardware developments, whereby complete humanoid robotic systems are as well described as partial solutions. In the second part diverse results around the biped motion of humanoid robots are presented. The autonomous, efficient and adaptive two-legged walking is one of the main challenge in humanoid robotics. The two-legged walking will enable humanoid robots to enter our environment without rearrangement. Developments in the field of visual sensors, data acquisition, processing and control are to be observed in third part of the book. In the fourth part some "mind building" and communication technologies are presented.



**InTech Europe**
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

**InTech China**
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821