Auton Robot
DOI 10.1007/s10514-011-9263-y

Two-handed gesture recognition and fusion with speech to command a robot

B. Burger · I. Ferrané · F. Lerasle · G. Infantes

Received: 6 February 2010 / Accepted: 28 November 2011

Abstract Assistance is currently a pivotal research area in robotics, with huge societal potential. Since assistant robots directly interact with people, finding natural and easy-to-use user interfaces is of fundamental importance. This paper describes a flexible multimodal interface based on speech and gesture modalities in order to control our mobile robot named Jido. The vision system uses a stereo head mounted on a pan-tilt unit and a bank of collaborative particle filters devoted to the upper human body extremities to track and recognize pointing/symbolic mono but also bi-manual gestures. Such framework constitutes our first contribution, as it is shown, to give proper handling of natural artifacts (self-occlusion, camera out of view field, hand deformation) when performing 3D gestures using one or the other hand even both. A speech recognition and understanding system based on the Julius engine is also developed and embedded in order to process deictic and anaphoric utterances. The second contribution deals with a probabilistic and multi-hypothesis interpreter framework to fuse results from speech and gesture components. Such interpreter is shown to improve the classification rates of multimodal commands compared to using either modality alone. Finally, we report on successful live experiments in human-centered settings. Results are reported in the context of an interactive manipulation task, where users specify local motion commands to Jido and perform safe object exchanges.

Keywords Human-robot interaction · Multiple object tracking · Two-handed gesture recognition · Vision and speech probabilistic fusion

Electronic supplementary material The online version of this article (doi:10.1007/s10514-011-9263-y) contains supplementary material, which is available to authorized users.

B. Burger (✉) · F. Lerasle
CNRS, LAAS, 7 avenue du Colonel Roche,
31077 Toulouse Cedex, France
e-mail: bburger@laas.fr

F. Lerasle
e-mail: lerasle@laas.fr

B. Burger · I. Ferrané
IRIT, Université de Toulouse, 118 route de Narbonne,
31062 Toulouse Cedex, France

I. Ferrané
e-mail: ferrane@irit.fr

I. Ferrané · F. Lerasle
Université de Toulouse, UPS, INSA, INP, ISAE; UT1, UTM,
LAAS, 31077 Toulouse Cedex, France

G. Infantes
Onera, 2 avenue Edouard Belin, 31055 Toulouse Cedex 4, France
e-mail: guillaume.infantes@onera.fr

1 Introduction

Having robots to assist people in human-centered environments is a goal that the robotics community has aspired to for many years (Fong et al. 2003). Such an assistant robot needs both spatial and transactional intelligence. Spatial intelligence, based on environment perception capabilities, means being able to understand and navigate in human-centered environment. Transactional intelligence, based on human perception capabilities, means being able to communicate meaningfully with a human user. In this paper, we will focus on this second kind of abilities and especially on peer to peer proximal interaction. In order to perform such an interaction, a robot has to be equipped with a multimodal user interface enabling to control the robot using several nat-

Published online: 21 December 2011

Fig. 1 Various H/R situations including multimodal commands

ural means like tactile senses,[1] speech and human body motion (*e.g.* handed gestures) senses which are here considered. Figure 1 illustrates some various human-robot situations including natural/multimodal commands the robot has to understand when using such interface.

The paper is organized as follows. Section 2 depicts an overview of related work and introduces our contributions. Section 3 describes the Jido platform and the multimodal interface in its software architecture. Section 4 presents the binocular tracking of the user's head and hands in order to interpret two-handed gestures, mainly symbolic or deictic, using Hidden Markov Models (HMMs). Section 5 includes a presentation of our system dedicated to verbal communication between our robot assistant and human users. Then, it depicts the multimodal interface based on the probabilistic fusion of speech and gesture interpretation results. Section 6 details live experiments. The aim is to characterize the whole interface during several trials of a human-robot interaction scenario. Lastly, Sect. 7 summarizes our contributions and discusses future extensions.

2 Overview and contributions

2.1 Related work

An increasing number of robotic systems are equipped for interaction with human users, each robot being designed for a specific objective. This specific design often limits their ability to perform complex Human-Robot Interaction (HRI). For example, guide robots in public places (see an overview in Arras and Burgard 2002) are intended to interact with several users, so they do not perform peer to peer HRI and usually do not need to recognize gestures. Other robotic platforms achieve more peer to peer interaction using a variety of sensory systems. Godot (Theobalt et al. 2002), Coyote (Skubic et al. 2004), Maggie (Gorostiza et al. 2006) and Dynamaid (Stückler et al. 2009), use speech recognition. CompanionAble (Badii and Thiemert 2009) also includes simple vision functions. Wakamaru (Harte and Jarvis 2007) uses a panoramic camera, Papero[2] and ALPHA (Bennewitz et al. 2005) detect faces, and Pearl (Pineau et al. 2003) performs face tracking. For these robots, focus has been on their appearance and their ability to communicate verbally in a natural way. Considering that 65% of the information in communication acts is nonverbal (Davis 1971) (namely motion), the ability for these robots to be understood and accepted by human users is limited.

Only few robots are able to perform more advanced interaction due to multimodal perception, combining speech recognition with gesture recognition. Gesture recognition is facilitated by the tracking of human limbs (usually hands or/and head). The robots named Biron (Maas et al. 2006) and ALBERT (Rogalla et al. 2004) benefit from such a multimodal user interface through 2D tracking and recognition of mono-manual pointing gestures. Pioneer (Yoshizaki et al. 2002) has similar skills but aims at recognizing more symbolic gestures; it tracks one human's hand in order to recognize the pattern currently painted. Cosero (Axenbeck et al. 2008) exhibits capabilities of two-handed gesture recognition but these are limited to the image plane. Even if these robots are able to perform advanced interaction with a human, 3D approaches are more suited to estimate human motions in depth, which naturally occurs when interacting with robots. The most advanced robotic systems (Hanafiah et al. 2004) and ARMAR (Stiefelhagen et al. 2004), are equipped for 3D tracking of the head and the two hands of their human user. The first one does not perform real gesture recognition, but assumes that speech and gestures are perfectly correlated *e.g.* if the user holds his hand some time in the same position while saying something, the robot considers that the human was pointing to an object. The second one recognizes mono-manual pointing gestures through HMMs, applying a relaxed strategy in terms of false positives (Nickel and Stiefelhagen 2006). This recognition step is triggered by speech. Both systems enable the user to communicate with the robot using gestures to point some objects out. Symbolic

[1] Tactile sense is out of the paper scope.

[2] See the http://fr.wikipedia.org/wiki/PaPeRo.

gestures are surprisingly not used: if these are less central than deictic ones, they are naturally used by humans in addition to speech (*e.g.* clapping hands while saying "Bravo", waving hands while saying "stop"). Recognizing such gestures means being potentially able to recognize two handed gestures. This requires to differentiate strongly the left hand from the right one during tracking.

A last observation concerns common underlying assumptions. Mono-manual hand gestures are usually presupposed *e.g.* Coyote (Skubic et al. 2004), ARMAR (Stiefelhagen et al. 2004), Pioneer (Yoshizaki et al. 2002), Robox (Siegwart 2003; Corradini and Gross 2000), and/or upper human body extremities are usually tracked separately *e.g.* RCB-1 (Park et al. 2005), Robovie (Hasanuzzaman et al. 2007), ARMAR (Nickel and Stiefelhagen 2006), inevitably inducing tracking failures when they overlap. To our best knowledge, the proper and simultaneous motion analysis of all the upper human body extremities has not yet been integrated on a mobile robot although a good gesture tracker is essential for any further gesture recognition process. This might open an increasing number of interaction possibilities, in particular through the use of the non-dominant hand in gesture recognition and/or two-handed gesture recognition.

Gesture recognition, possibly combined with speech for multimodal communication, has recently received attention in the HRI community. Visual gestures show human thoughts, replays, complements, accents, and adjust verbal information. Therefore, vision-based gesture interpretation is valuable in environments where speech-based communication may be garbled or drowned out. Moreover, the mutual assistance between the robot's speech and vision capabilities enables a user to robustly specify location references in verbal statements. Combined with pointing gestures, such prominent commands open up the possibility of intuitively indicating objects and locations *e.g.* to make the robot change its direction/position or to mark objects. Nevertheless, it can be argued that, in these cases, vision techniques for human perception and natural language processing have mostly been studied as two separate research topics (Prodanov and Drygajlo 2003a; Skubic et al. 2004; Triesch and Von der Malsburg 2001; Waldherr et al. 2000), rather than been combined.

A lot of studies aim to couple these two communication channels (audio and video) and several robots are now equipped with such a multimodal interface, but in all of them speech remains the main channel. The easiest strategy was developed by Hanafiah et al. (2004) which considers that speech and gestures are perfectly correlated, but does not perform real gesture recognition. Conversely, in Yoshizaki et al. (2002), vision is only used if a need is detected by speech, leading to unnatural interaction. Rogalla et al. (2004) fuses events from these channels to define the action to be performed by the robot, but the system is handicapped by limited visual capabilities (simplistic tracking, only 2D gestures). Finally, the most advanced multimodal interface is probably (Stiefelhagen et al. 2004) which fuses speech and mono-manual 3D gestures in a probabilistic way, but evaluations of the whole robotic system are not mentioned.

2.2 Contributions

The first contribution in this paper concerns the design of a real-time and robust tracking framework based on multiple and interactive particle filters in order to analyze the 3D motions of all the upper human body extremities from the onboard stereo head of a mobile robot. Using this multiple object tracker (MOT), enhanced recognition performances for mono and bi-manual hand 3D gestures are achievable. To our best knowledge, no system based on two-handed gesture recognition has been developed and integrated for interactive robots. The second contribution deals with a probabilistic and multi-hypothesis interpreter framework to fuse results from speech and gesture components. Guided by this multimodal fusion, the interpretation of deictic and symbolic actions can be improved compared to the results using solely speech or gesture. A last contribution concerns the integration of this interface into our mobile robot Jido as few existing studies have addressed onboard multimodal interfaces that can cope with both robotic and natural settings (Hanafiah et al. 2004; Gorostiza et al. 2006; Maas et al. 2006; Rogalla et al. 2004). The target scenario we address is a peer to peer HRI in which a human can ask the robot to move according to his/her command to mark or bring certain objects, etc. This application serves as a motivation for the general multimodal communication required for any mobile robot acting as an assistant.

3 The Jido platform and its software architecture

Our multimodal interface is embedded on a robot companion named Jido (Fig. 2) which consists of: a 6-DOF arm equipped with a videre stereo bank, a pan-tilt stereo system at the top of a mast and a laser range finder (LRF) in front. The embedded functionalities are managed by the `LAAS` layered software architecture (Fig. 3) and detailed in Alami et al. (1998).

Such functionalities enable Jido to:

1. Build maps and navigate in indoor environments thanks to the LRF sensor. The embedded functionalities are under the Base motion box (Fig. 3).
2. Recognize and manipulate objects thanks to the videre stereo system. A standard procedure consists in extracting the 3D position of the object using blob detection, then computing an arm trajectory which is executed by the dedicated module, all this is done within the Object Recognition and Manipulation boxes.

3. Perceive humans (detailed module) using the pan-tilt stereo system on the mast, namely (i) detection/recognition and view-based tracking from the `Face recognition` modules, (ii) control the pan-tilt unit mounted stereo head from the `Human Position` module, (iii) 3D gestures tracking and recognition from the `GEST` modules, (iv) speech utterance recognition and interpretation from the `RECO` module, (v) fusion of speech and gesture modalities in the `FUSION` module.
4. Talk to the human user using the Speech synthesis module.

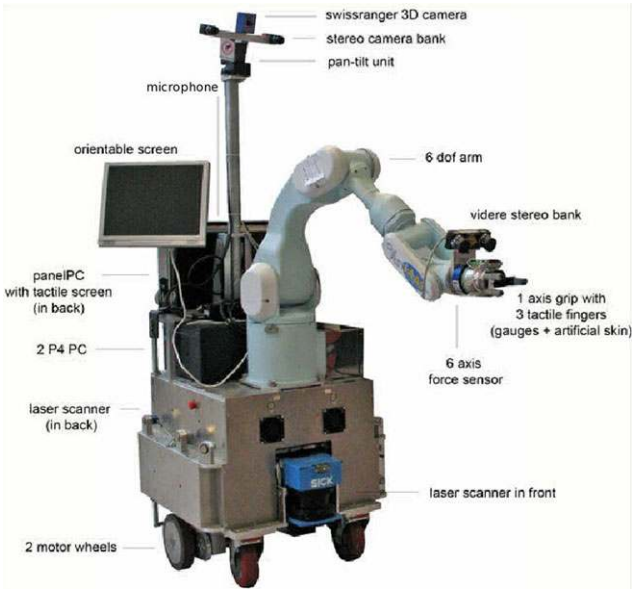

Fig. 2 The robot Jido

The following sections detail the modules dedicated to human perception with a special emphasis on `GEST` (Sect. 4) and `FUSION` (Sect. 5) which requires `RECO` (Sect. 5.1). These modules constitute our main contributions.

4 Visual perception of the robot user

In this section, we describe the setup and approach for tracking and recognizing dynamic gestures in the video stream. We propose a new framework based on multiple interactive particle filters dedicated to the upper human body extremities (head and both hands). Then, we use Hidden Markov Models (HMM) applied to temporal sequences of hand poses for the occurrence of a mono or bi-manual hand gesture. We present here below a brief survey of the literature in the area of gesture tracking and put our tracker in perspective.

4.1 Overview and related work

Visual tracking of human body parts has been studied extensively over the past few years. The reader can refer to two recent and comprehensive surveys (Erol et al. 2007; Murphy-Chutorian and Trivedi 2008) related to head and hand tracking in the Human-Computer Interaction (HCI) literature. A main problem hampering most of the approaches surveyed is that they rely on underlying assumptions that may be suitable for HCI applications but not in HRI ones: obtrusive sensors (Fels and Hinton 1997), static and high contrast background (Huang et al. 2002; Isard and Blake

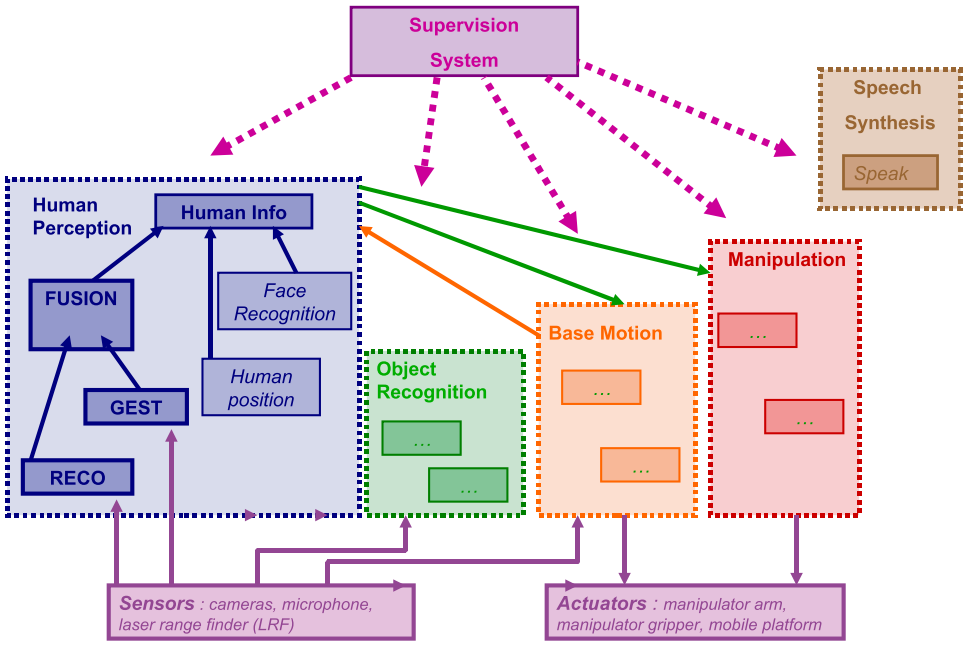

Fig. 3 Jido's software architecture

1998a; Just et al. 2004), specific clothing appearance (Azad et al. 2007; Waldherr et al. 2000), etc. The aforementioned surveys highlight that particle filters have become increasingly popular, especially for HRI applications. Their popularity stems from their ability to: (i) deal with multimodal probability distributions, (ii) fuse diverse kinds of measurements easily in a probabilistically principled way, allowing them to handle clutter (Pérez et al. 2004).

The approaches surveyed can be split in two broad categories. The former relies on view-based models *i.e.* mono (Chen et al. 2003; Rogalla et al. 2004; Thayananthan et al. 2003), or bi-manual (Hasanuzzaman et al. 2004; Park et al. 2005) gestures analyzed in the image space. These 2D approaches are inappropriate for estimating motion in depth, and so have difficulty interpreting natural gestures, especially deictic ones, which occur in a 3D space.

The second category, based on 3D generative models, involves the best match between model projections and image, and recovering the associated 3D human posture. The conventional approach attempts to infer the 3D global pose and all the upper human body joint angles in the video stream (see a survey in Moeslund et al. 2006). These approaches require modeling detailed human geometry and so are not generally person independent. Moreover, particle filters (PF) for full DOF tracking have proven to be computationally expensive precluding real-time implementation and keeping such an approach from integration on mobile autonomous robots. Finally, it can be argued that the full reconstruction of the entire kinematic chain is not essential for gesture recognition. In the vein of Bernier and Collobert (2001), Nickel and Stiefelhagen (2006), we have sought an alternative 3D representation of the human body. We model solely the upper human body extremities with deformable and coarse ellipsoids for computational tractability reasons. Compared to conventional particle filters for full DOF tracking of the upper human body parts (Fontmarty et al. 2007), our sparse tracker is shown to limit the computational complexity. Moreover, our deformable models allow us to deal with the 3D orientation of the hands and their heavy deformations like the opening/closing actions of the palm.[3] Finally, through the addition of flexible geometric constraints between these ellipsoids, this representation is person independent.

Given this sparse 3D model of the upper human body extremities, our multiple object tracker (MOT) framework stands out from the literature due to the following two enhancements.

The first improvement concerns interactive mechanisms to ensure the consistency both in 2D and 3D spaces. To date, MOTs have been devoted to the tracking of multiple persons *e.g.* Qu et al. (2007), Zhao and Nevatia (2004) but rarely to multiple parts of a single person (Bernier and Collobert 2001; Nickel and Stiefelhagen 2006).

A key problem remains the well-known "coalescence" phenomenon when targets undergo partial or complete self-occlusion. In other words, MOT often associates more than one human body part trajectories to some targets while loses track for others. Two main classes of MOT, with their respective advantages and drawbacks, can be considered. The former, widely accepted in the computer vision community, exploits a single joint state representation which concatenates all of the targets' states together, and the latter uses distributed filters, namely one filter per target. The "coalescence" problem might be more correctly handled during the joint inference underlying centralized approaches. However, they are not scalable due to their nature of exponential complexity. For example, JPDAF[4]-based methods like Bar-Shalom and Jaffer (1998), Rasmussen and Hager (2001) suffer from the combinatorial complexity due to the exhaustive enumeration for data association; and sampling-based stochastic approaches like Isard and Blake (2001), Zhao and Nevatia (2004) are confronted by the exponential demand of the increase of particles with the state space dimensionality. Besides, usual distributed/decentralized approach *e.g.* based on multiple independent particle filters (MIPF), suffers from this "coalescence" phenomenon. More recent investigations (Qu et al. 2007; Yu and Wu 2004) have highlighted that interactively distributed MOT (called IDMOT) limit such phenomena as each tracker, devoted to a given target, relies on the other neighboring targets' status.

A last remark concerns 3D two-handed gestures which have received for now little attention in the literature. The few 3D gesture trackers both in HCI (Bernier and Collobert 2001) and HRI (Nickel and Stiefelhagen 2006) propose no advanced mechanisms to properly handle self-occlusions and hand-hand occlusion which usually appear when performing two-handed gestures in 3D. Clearly, tracking two-handed gestures is a requisite for HRI application as humans should be able to operate the robot with either the dominant or non-dominant hand, and some commanding gestures might be naturally mono or bi-manual *e.g. "hello", "go forward", etc*.

The second improvement concerns the automatic (re)-initialization. The CONDENSATION—for "Conditional Density Propagation" (Isard and Blake 1998b)—is the most popular PF strategy such that the particles are drawn according to a proposal distribution based solely on the system dynamics and so 'blindly" w.r.t. the measurement. In practice, tracking soon goes astray if no recovery/initialization process is added. We propose an extension of the IDMOT framework (called IIDMOT), in the vein of the ICONDENSATION strategy (Isard and Blake 1998a), which provides

[3] Free-form hand gestures are clearly the most natural nonverbal means of communication.

[4] For Joint Probabilistic Data Association Filter.

the capability to recover the targeted limbs after temporary target loss, camera out of sight, etc. Our principle consists in also sampling some particles from the dynamics and some w.r.t. visual detectors for online re-initialization. As pointed out by Pérez et al. (2004), we are convinced that a crucial design issue in PF is the choice of the proposal distribution.

We propose comparative evaluations demonstrating that our IIDMOT approach outperforms the conventional MIPF *i.e.* independent filters, and IDMOT *i.e.* particle filters with interaction mechanisms but without (re)-initialization capabilities. Such tracking adaptations are expected to fulfill all the robotic requirements stated above, and to have immediate impact on gesture recognition in a bottom-up strategy.

Regarding this gesture recognition stage, our approach based on HMMs is more or less conventional even if the existing systems embedded on mobile robots and with automatic (re)-initialization capability are rather rare. The adaptations are twofold. First, our approach, by handling the time series of the two hand positions, stands out from most of the existing 3D approaches in the HRI literature (Corradini and Gross 2000; Park et al. 2005; Shimizu et al. 2006; Stiefelhagen et al. 2004; Triesch and Von der Malsburg 2001; Yoshizaki et al. 2002) which assume a dominant-handed gesture *i.e.* the right one. We focus on 3D mono but also bi-manual gestures as body motions occur in space while two-handed gestures have strong expression capabilities. The most similar approach to ours is that of Just et al. (2004) who have shown that incorporating two-hand movement information helps to improve the recognition of 3D mono or bi-manual gestures but only in the HCI context. Second, the hand 3D coordinates are transformed into the head-centered 3D coordinate system in order to become invariant regarding the location of the person, the pan-tilt unit and robot motions.

4.2 3D tracking of head and hands

Our IIDMOT framework is depicted in Table 1. Recall that particle filters aim to recursively approximate the posterior probability density function (pdf) $p(\mathbf{x}_t^i|z_{1:t})$ of the state vector $\mathbf{x}_t^i$ for body part i at time t given the set of measurements $z_{1:t}$ (given in (1)). A linear point mass combination

$$p(\mathbf{x}_t^i|z_{1:t}) \simeq \sum_{n=1}^{N} \omega_t^{i,n}.\delta(\mathbf{x}_t^i - \mathbf{x}_t^{i,n}), \quad \sum_{n=1}^{N} \omega_t^{i,n} = 1, \tag{1}$$

is determined which expresses the selection of a value—or "particle"—$\mathbf{x}_t^{i,n}$ for target i at time t with probability—or "weight"—$\omega_t^{i,n}$. $\delta(.)$ is the Dirac delta distribution. An approximation of the conditional expectation of any function of $\mathbf{x}_t^i$, such as the minimum mean square error (MMSE) estimate $E_{p(\mathbf{x}_t^i|z_{1:t})}[\mathbf{x}_t^i]$, then follows (step 5, Table 1).

In our framework, when two particles $\mathbf{x}_t^{i,n}$ and $\mathbf{x}_t^{j,n}$ for target i and j do not interact one with the other, *i.e.* their relative Euclidean distance exceeds a predefined threshold (noted d_{TH}, step 7 in Table 1), the approach performs like multiple independent trackers. When they are in close proximity, magnetic repulsion and inertia likelihoods are added in each filter to handle the aforementioned problems. Following (Qu et al. 2007), the repulsion "weight" $\varphi_1(.)$ follows (in (2))

$$\varphi_1(\mathbf{x}_t^{i,n}(z_t^i), \mathbf{x}_t^{j,n}(z_t^j)) \propto 1 - \frac{1}{\beta_1} \exp\left(-\frac{D_{i,n}^2}{\sigma_1^2}\right), \tag{2}$$

with β_1 and σ_1 two normalization terms being determined *a priori*. $D_{i,n}$ denotes the Euclidean distance between particle $\mathbf{x}_t^{i,n}$ and particle $\mathbf{x}_{t,k}^{j}$ at iteration k. Practically, if the analyzed tracker i is isolated from target j, it will only implement MIPF to reduce the computational costs. When it becomes closer or interacts with tracker j (inducing occlusions), it will activate the iterative IDMOT (steps 7–19, Table 1) to handle the "coalescence" problem. The principle can be extended to a 3-clique $\{z^i\}_{i=1,2,3}$. The inertia "weight" $\varphi_2(.)$ considers the target's motion vector $\overrightarrow{v_1}$ from the states in previous two frames in order to predict its motion vector $\overrightarrow{v_2}$ for the current. The function then follows

$$\begin{aligned}\varphi_2(\mathbf{x}_t^{i,n}, \mathbf{x}_{t-1}^{i,n}, \mathbf{x}_{t-2}^{i,n}) \\ \propto 1 + \frac{1}{\beta_2} \exp\left[-\frac{(\|\overrightarrow{v_1}\| - \|\overrightarrow{v_2}\|)^2}{\sigma_{22}^2}\right] \\ \times \exp\left(-\frac{\theta_{i,n}^2}{\sigma_{21}^2} \cdot \frac{\|\overrightarrow{v_1}\|^2}{\sigma_{22}^2}\right),\end{aligned} \tag{3}$$

with β_2 a normalization term. $\theta_{i,n}$ represents the angle between the above vectors while σ_{21} and σ_{22} characterize the variance of motion vector direction and speed.

Our IDMOT particle filter, named IIDMOT, follows this principle but is extended in three ways. First, the conventional CONDENSATION (Isard and Blake 1998b) strategy is replaced by the ICONDENSATION (Isard and Blake 1998a) one whose importance function $q(.)$ in step 3 of Table 1 permits automatic (re)-initialization when the targeted human body parts appear or re-appear in the scene. Thus, the classical importance function $q(.)$ (in (4)) based on dynamics $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and the prior p_0 can be extended to consider measurements z_t in the sub function $\pi(.)$ so that, with $\alpha \in [0; 1]$,

$$q(\mathbf{x}_t^{i,n}|\mathbf{x}_{t-1}^{i,n}, z_t^i) = \alpha\pi(\mathbf{x}_t^{i,n}|z_t^i) + (1-\alpha)p(\mathbf{x}_t^{i,n}|\mathbf{x}_{t-1}^{i,n}). \tag{4}$$

The prominent new term $\pi(.)$ aims to sample a particle subset according to visual detectors *i.e.* skin colored blob

Table 1 Our IIDMOT algorithm

1: **IF** $t=0$, **THEN** Draw $\mathbf{x}_0^{i,1},\ldots,\mathbf{x}_0^{i,j},\ldots,\mathbf{x}_0^{i,N}$ i.i.d. according to $p(\mathbf{x}_0^i)$, and set $w_0^{i,n}=\frac{1}{N}$ **END IF**
2: **IF** $t\geq 1$ **THEN** $\{$—$[\{\mathbf{x}_{t-1}^{i,n},w_{t-1}^{i,n}\}]_{n=1}^{N}$ being a particle description of $p(\mathbf{x}_{t-1}^i|z_{1:t-1}^i)$—$\}$
3: "Propagate" the particle $\{\mathbf{x}_{t-1}^{i,n}\}_{n=1}^{N}$ by independently sampling $\mathbf{x}_t^{i,n}\sim q(\mathbf{x}_t^i|\mathbf{x}_{t-1}^{i,n},z_t^i)$
4: Update the weight $\{w_t^{i,n}\}_{n=1}^{N}$ associated to $\{\mathbf{x}_t^{i,n}\}_{n=1}^{N}$ according to the formula $w_t^{i,n}\propto w_{t-1}^{i,n}\frac{p(z_t^{i,c,n},z_t^{i,s,n}|\mathbf{x}_t^{i,n}).p(\mathbf{x}_t^{i,n}|\mathbf{x}_{t-1}^{i,n})}{q(\mathbf{x}_t^{i,n}|\mathbf{x}_{t-1}^{i,n},z_t^i)}$, prior to a normalization step so that $\sum_{n=1}^{N}w_t^{i,n}=1$
5: Compute the conditional mean of any function of $\hat{x}_t^i$, *e.g.* the MMSE estimate $\mathrm{E}_{p(\mathbf{x}_t^i|z_{1:t}^i)}[\mathbf{x}_t^i]$, from the approximation $\sum_{n=1}^{N}w_t^{i,n}\delta(\mathbf{x}_t^i-\mathbf{x}_t^{i,n})$ of the posterior $p(\mathbf{x}_t^i|z_{1:t}^i)$
6: **FOR** $j=1:i$, **DO**
7: **IF** $d_{ij}(\hat{\mathbf{x}}_{t,k}^i,\hat{\mathbf{x}}_{t,k}^j)<d_{TH}$ **THEN**
8: Save link(i,j)
9: **FOR** $k=1:K$ iterations, **DO**
10: Compute φ_1,φ_2
11: Reweight $w_t^{i,n}=w_t^{i,n}.\varphi_1.\varphi_2$
12: Normalization step for $\{w_t^{i,n}\}_{n=1}^{N}$
13: Compute the MMSE estimate $\hat{\mathbf{x}}_t^i$
14: Compute φ_1,φ_2
15: Reweight $w_t^{j,n}=w_t^{j,n}.\varphi_1.\varphi_2$
16: Normalization step for $\{w_t^{j,n}\}_{n=1}^{N}$
17: Compute the MMSE estimate $\hat{\mathbf{x}}_t^j$
18: **END FOR**
19: **END IF**
20: **END FOR**
21: At any time or depending on an "efficiency" criterion, resample the description $[\{\mathbf{x}_t^{i,n},w_t^{i,n}\}]_{n=1}^{N}$ of $p(\mathbf{x}_t^i|z_{1:t}^i)$ into the equivalent evenly weighted particles set $[\{\mathbf{x}_t^{(s^{i,n})},\frac{1}{N}\}]_{n=1}^{N}$, by sampling in $\{1,\ldots,N\}$ the indexes $s^{i,1},\ldots,s^{i,N}$ according to $P(s^{i,n}=j)=w_t^{i,j}$; set $\mathbf{x}_t^{i,n}$ and $w_t^{i,n}$ with $\mathbf{x}_t^{(s^{i,n})}$ and $\frac{1}{N}$
22: **END IF**

detection for hands/head (Just et al. 2004) and frontal face detection (Viola and Jones 2001) which, despite their sporadicity, are very discriminant when present. Practically, $\alpha\%$ of the particles are drawn according to 3D ellipsoids after triangulation on image ROIs corresponding to detected skin blobs or faces. Secondly, the IDMOT particle filter, pioneered in Qu et al. (2007) for image-based tracking of multiple objects without (re)-initialization capabilities, is here extended to estimate the 3D pose of multiple deformable body parts of a single person. The third line of investigation concerns data fusion, as our observation model is based on a robust and probabilistically motivated integration of multiple cues. Fusing 3D and 2D (image-based) information from the video stream of a stereo head, with cameras mounted on a mobile robot, allows us to benefit both from reconstruction-based and appearance-based approaches.

Our novel IIDMOT strategy combines the advantages of ICONDENSATION and IDMOT in order to jointly handle the (re)-initialization and "coalescence" problems. It aims to fit the projections of a sphere and two deformable ellipsoids (respectively representing the head and the two hands) throughout the video stream, using the estimation of the 3D location $\mathcal{X}=(X,Y,Z)'$, the orientation $\Theta=(\theta_x,\theta_y,\theta_z)'$, and the axis length[5] $\Sigma=(\sigma_x,\sigma_y,\sigma_z)'$ for ellipsoids. All these parameters are accounted for in the state vector $\mathbf{x}_t^i$ related to target i for the t-th frame. With regard to the dynamics model $p(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i)$, the 3D motions of observed gestures are difficult to characterize over time. This weak knowledge is formalized by defining the state vector as $\mathbf{x}_t^i=[\mathcal{X}_t,\Theta_t,\Sigma_t]'$ for each hand and assuming that its entries evolve according to mutually independent random walk models, viz. $p(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i)=\mathcal{N}(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i,\Lambda)$, where $\mathcal{N}(.|\mu,\Lambda)$ is a Gaussian distribution in 3D with mean μ and covariance Λ being determined *a priori*. Our importance function $q(.)$ followed by our multiple cues based measurement function $p(z_t^i|\mathbf{x}_t^i)$ are depicted below. Recall that α percent of the particles are sampled from detector $\pi(.)$ ((4)). These are also drawn from Gaussian distribution for head or hand configuration but deduced from skin color blob segmentation in the CIE-Lab color space. The centroid and associated covariances of the matched regions are finally triangulated using the parameters of the calibrated stereo setup. For the weight updating step (step 4, Table 1), each ellipsoid defined by its configuration $\mathbf{x}_t^i$ is then projected in one of the two image planes. Given $Q=\left[\begin{smallmatrix}A & b\\ b' & c\end{smallmatrix}\right]$ the associated 4×4 symmetric matrix, the set of image points $\mathbf{x}$ that belongs to the projection contours verify the following expression: $\mathbf{x}'.(\mathbf{b}\mathbf{b}'-c\mathbf{A}).\mathbf{x}=0$.

The measurement function fuses skin color information but also motion and shape cues. Each ellipsoid for any pose $\mathbf{x}_t^i$ leads to an ellipse in the image plane after projection.

[5] To take into account the hand orientation in 3D.

The pixels are then partitioned into a set of target pixels O belonging to this ellipse, and B a set of pixels which are assumed not corresponding to both hands or head. Assuming pixel-wise independence, the skin color-based likelihood is factored as

$$p(z_t^{i,c}|\mathbf{x}_t^i) = \prod_{o\in O} p_s(o|\mathbf{x}_t^i) \prod_{b\in B}[1 - p_s(b|\mathbf{x}_t^i)], \tag{5}$$

where $p_s(j|\mathbf{x}_t^i)$ is the skin color probability at pixel location j given $\mathbf{x}_t^i$. Using only color cue for the model to image fitting is not sufficiently discriminant in our robotics context. We also consider a likelihood $p(z_t^{i,s}|\mathbf{x}_t^i)$ which combines motion and shape cues. In some H/R situations, it is highly possible that the targeted limbs are moving, at least intermittently. We thus use the cost function (6) which is based on two terms: the penalty term favours the moving edges without removing the static ones (see the first term). The result is that a tracker based on this will prefer to stick with the moving edges, but in their absence its fallback is to use the static ones.

$$p(z_t^{i,s}|\mathbf{x}_t^i) \propto \exp\left(-D^2/2\sigma_s^2\right),$$

$$D = \sum_{j=1}^{N_p} |x(j) - z(j)| + \rho\gamma(z(j)). \tag{6}$$

This equation depends on the sum of the squared distances between N_p points uniformly distributed along the ellipsoid contours $\mathbf{x}$ and their nearest image edges z. σ_s is a standard deviation being determined *a priori*. Given $\overrightarrow{f}(z_t(j))$ the optical flow vector at pixel $z(j)$, $\gamma(z(j)) = 0$ (resp. 1) if $\overrightarrow{f}(z(j)) \neq 0$ (resp. if $\overrightarrow{f}(z(j)) = 0$) and $\rho > 0$ terms a penalty.

Finally, assuming the cues to be mutually independent, the unified measurement function in step 4 (Table 1) is formulated as

$$p(z_t^{i,c}, z_t^{i,s}|\mathbf{x}_t^i) = p(z_t^{i,c}|\mathbf{x}_t^i).p(z_t^{i,s}|\mathbf{x}_t^i). \tag{7}$$

4.3 Gesture recognition

There are two problems to address when dealing with dynamic gesture recognition: spotting and classification. On one hand, spotting aims at identifying the beginning and/or the end of a gesture given a continuous stream of data which is made up of random sequences of legitimate gestures and non-gestures (moves in between meaningful gestures). All the gestures are here assumed to start and end in the same natural/rest position (the hands hanging along the body). On the other hand, given an isolated gesture sequence, classification outputs the class the gesture belongs to, among a vocabulary composed of 5 (resp. 7) deictic (resp. symbolic) gestures (see Table 4) whose choices were motivated by the particular HRI scenario described in Sect. 6. The 7 symbolic gestures, defined by their motion templates, are namely: *"greetings" (with one or two hands), "introducing oneself", "come to me" (with one or two hands), "stop", "go away"*. Besides, the 5 deictic gestures depend on the coarse pointed direction *i.e. "ahead", "bottom left", "bottom right", "top left", "top right"*.

For deictic gestures, the pointing direction is calculated by the connecting line between the center of the head and the hand assuming this rough direction is enough to distinguish between sparse objects in the human vicinity. In order to make our gesture recognition independent of the position of the user in relation to the robot, we define our vector in a coordinate system centered on the head whose y and z axis constitute the 'human plane' *i.e.* the plane formed by the head and both hands of the human at his rest position. Given the outputs of our IIDMOT multi tracker, all models are trained by means of the EM-algorithm using the following 9-dimensional feature vector derived from the tracked positions of both head and hands

$$\mathbf{x}_k = (D_{H-Lh}, \Theta^L, D_{H-Rh}, \Theta^R, D_{Lh-Rh})',$$

where D_{H-Lh} is the distance between the head and the left hand location in space, D_{H-Rh} is the analogous term for the right hand. $\Theta^i = (\theta_x^i, \theta_y^i, \theta_z^i)$ is the orientation of hand i with regard to the 'human plane' and D_{Lh-Rh} is the distance between the two hands.

Unlike Stiefelhagen et al. (2004), each complete gesture is here straightforwardly modeled by a dedicated HMM. The topology of the HMMs is determined empirically *i.e.* five state models were found to be the best compromise between performance and computational cost. We use discrete HMMs whose space size (number of clusters per variable) and geometry (size of each of these clusters) are determined through a self-organizing map, or Kohonen network. These HMM models are trained using the Expectation-Maximization algorithm. The reader can here refer to Fox et al. (2006) for more details about our HMM tuning.

4.4 Off-line experiments

Prior to their integration on our mobile robot, experiments on a data set of 10 sequences (1214 stereo images) acquired from the robot are performed off-line in order to: (i) determine the optimal parameter values of our strategy, and (ii) characterize its performance. These gestures are performed either using the left or the right hand or both. Moreover, this sequence set involves variable viewing conditions, namely illumination changes, clutter, self-occlusions or out of field of view. Figure 4 shows snapshots of two sequences recorded while performing gestures involving self-occlusion like *"greetings"* and *"go to my left"*. For each frame, the

Fig. 4 Two sequences involving occlusion (for gestures *"greetings" (with one hand)* (**a**) and *"go to my left"* (**b**). Each is run with simple MOT (*top*) and our IIDMOT (*bottom*). More gesture examples can be found at http://brice.burger.pagesperso-orange.fr/these/gestes/index.html

template depicts the projection of the MMSE estimate for each ellipsoid. The IIDMOT strategy, by using magnetic repulsion and inertia likelihoods enables to handle the "coalescence" problem. Furthermore by drawing some particles according to the detector output, this strategy permits automatic re-initialization and aids recovery after loss of observability. Moreover, given the estimated state vector, both motion and shape parameters for dynamic gesture recognition are obtained.

Quantitative performance evaluation has been carried out on the sequence set. Since the main concern of tracking is the correctness of the tracker results, location as well as label, we compare the tracking performance quantitatively by defining the false position rate (FR_p) and the false label rate (FR_l). As we have no ground truth, failure situations must be defined. Failing to associate a tracker with one of the targets in (at least) one image plane will correspond to a position failure, while a tracker associated with the wrong target will correspond to a label failure. Table 2 presents the performances using multiple independent particle filters (MIPF) (Isard and Blake 1998b), conventional IDMOT (Qu et al. 2007) strategy, and our IIDMOT strategy with data fusion (see Sect. 4.2). Our IIDMOT strategy is shown to outperform the conventional approaches with a slight increase in computation time. The MIPF strategy suffers especially from "coalescence" problems due to a lack of interaction modeling between trackers while the IDMOT strategy does not recover the target after transient loss. These results have been obtained for the "optimal" tracker parameter values listed in Table 3.

Given the IIDMOT outputs, we evaluated (also off-line) our gesture recognizer using a data set of 324 image sequences pre-acquired from the robot Jido. These sequences only contain meaningful gestures. The data set is split into two subsets: the training set T (2/3 of the data-base) and the test set T_e (1/3 of the data-base).

Table 4 reports on quantitative evaluations in terms of recognition rate for the overall test sequences. Symbolic

Table 2 Quantitative performance and speed comparisons

Method	MIPF	IDMOT	IIDMOT
FR_p	29%	18%	4%
FR_l	9%	1%	1%
Speed (fps)	15	12	10

gestures like "greetings" and "introducing oneself" (user pointing himself when saying his name) are well recognized. Nevertheless confusion rates are quite hight between some gestures fairly close to each other. One time upon three, "pointing top right" is recognized as "greetings" using one hand, which is not the case of "pointing top left", users mainly waving their right hand when greeting. There is also confusion between "pointing bottom right" and "pointing ahead" and vice versa, which is understandable if the right hand is used when pointing ahead. Some other confusion cases seem more difficult to explain because they results from multiple converging facts. For example, "pointing top-left" and "bottom-left" can be performed just above or below the horizontal plane, which can be, with an estimation error from the tracker, very close. In these experiments, 72% of the gestures performed were correctly classified. This rate is comparable to the 70% obtained in Just et al. (2004) for mono or bi-manual 3D gesture recognition but in the HCI context. Here, the most prominent error was a failure to recognize *"come to me (one hand)"*, which we attribute to a poor set of motion templates for this gesture. We also observe that bi-manual gestures are very well classified compared to their mono-manual counterparts (see *"greetings"* and *"come to me"* gestures).

Speech being the other modality we consider in our multimodal interaction scenario, the following section describes how gesture and speech are combined through a fusion step.

Table 3 Parameter values used in our IIDMOT tracker

Symbol	Meaning	Value
N	number of particles per filter	100
α	coeff. in the importance function $q(\mathbf{x}_t^{i,n}\|\mathbf{x}_{t-1}^{i,n}, z_t^i)$	0.4
K	number of iterations in PF algorithm	4
d_{TH}	Euclidean distance between particles in PF algorithm	0.5
–	image resolution	256×192
–	color-space for skin-color segmentation	CIE Lab
N_p	number of points along the ellipsoid contours	20
σ_s	standard in likelihood $p(z_t^{i,s}\|\mathbf{x}_t^i)$	36
ρ	penalty in likelihood $p(z_t^{i,s}\|\mathbf{x}_t^i)$	0.12
(σ_1, β_1)	coeff. in the repulsion "weight" φ_1	(0.12, 1.33)
$(\sigma_{21}, \sigma_{22}, \beta_2)$	coeff. in the inertia "weight" φ_2	(1.57, 0.2, 2.0)
Λ	standard deviation in random walk models	$\begin{pmatrix} 0.07 & 0.07 & 0.07 \\ 0.03 & 0.03 & 0.03 \\ 0.17 & 0.17 & 0.17 \end{pmatrix}$

Table 4 Recognition results for our gesture recognition system (in %)

Gestures given		Sensibility	Gestures recognized											
			1	2	3	4	5	6	7	8	9	10	11	12
selectivity		72	80	100	52	63	78	64	78	63	60	100	70	87
"come to me (one hand)"	(1)	36	36	0	18	0	0	0	0	18	18	0	0	9
"come to me (two hands)"	(2)	72	0	72	27	0	0	0	0	0	0	0	0	0
"pointing bottom left"	(3)	81	0	0	81	18	0	0	0	0	0	0	0	0
"pointing top left "	(4)	63	0	0	9	63	27	0	0	0	0	0	0	0
"greetings (two hands)"	(5)	100	0	0	0	0	100	0	0	0	0	0	0	0
"greetings (one hands)"	(6)	100	0	0	0	0	0	100	0	0	0	0	0	0
"introducing oneself"	(7)	100	0	0	0	0	0	0	100	0	0	0	0	0
"go away"	(8)	63	0	0	0	0	0	9	27	63	0	0	0	0
"pointing bottom right"	(9)	54	0	0	0	18	0	0	0	0	54	0	27	0
"pointing top right"	(10)	63	0	0	0	0	0	36	0	0	0	63	0	0
"pointing ahead"	(11)	63	0	0	18	0	0	0	0	0	18	0	63	0
"stop"	(12)	63	9	0	0	0	0	9	0	18	0	0	0	63

5 Multimodal fusion

In a natural interaction situation, a human user will readily use both channels, particularly in an object manipulation task or to emphasize part of his message. In such cases, these modalities are highly interconnected. Building a semantic representation of the user's message requires merging pieces of information extracted from the audio and video streams. In our framework, results from the audio components are provided by a speech recognition and interpretation step.

5.1 Embedded speech recognition and interpretation

The first aim of our work is to make an autonomous robotic platform able to process user utterances when performed in a natural communication situation. The autonomy issue led us to choose an embedded solution for speech recognition and to make it light enough in terms of computational and memory resources, since other processes are running on the same platform at the same time. Speech recognition in the HRI domain can go from recognizing a small set of keywords like in Prodanov and Drygajlo (2003b) where the robot takes initiative by asking the user only yes/no questions, to word sequence recognition in continuous speech. Recognition is generally performed using either a commercial speech recognition software (like Hermes in Bischoff and Graefe 2004, Maggie Gorostiza et al. 2006) or an open source speech recognition engine (like Stiefelhagen et al. 2004; Lee et al. 2001). In the former case, a server is gener-

ally dedicated to speech processing which is not performed by the robotic platform itself. In the latter case, as linguistic resources have to be provided, they can be designed to either reduce the computation cost or be adapted to the environment, for example in new acoustic conditions and interaction contexts. The module called RECO, which performs speech recognition and interpretation, is shortly described in the two following subsections.

5.1.1 Recognizing speech with RECO

Based on the Julius open source speech recognition engine (Lee et al. 2001) RECO uses a light version of acoustic and phonetic models in order to limit the computation cost. These 39 monophone models (37 for phonemes and 2 for short and long pauses) are HMM-based (3-state models with 32 Gaussians per state) and trained on 31 hours of Broadcast News recorded on French radios for a completely different task, namely Rich Transcription of French Broadcast News in the evaluation campaign ESTER (Galliano et al. 2005). Our application field being more task-oriented, the lexicon and the language model were specifically designed in order to take different types of user requests into account. These requests are ranging from simple commands to more complex queries, leading to spatial reference resolution (for location and object) because of their deictic or anaphoric aspect, as shown in Table 5. A medium size lexicon (246 words/428 pronunciations) was drawn up from the French lexical database BDLEX (Pérennou and de Calmès 2000) and context free grammars were designed to cover the different types of user requests, including some language flexibility. Using statistical language models like N-grams would have required a training step and a significant corpus of written or transcribed sentences which was not available. So, considering our task-oriented context, word sequences were described by sets of rules from which a set of 2334 different well-formed sentences can be derived. After phonetic and lexical decoding, the word sequence matching the syntactic rules given by the grammars with the best score (maximum likelihood) is proposed as the best speech recognition hypothesis. Hypotheses with lower scores generally differ from the best one, from a phonetic, lexical or syntactic point of view. For example, speech recognition could propose *"Hello, it is me Paul"* and *"Hello, it is Mickael"* as two different hypotheses of a same utterance.

To evaluate these results, and later those of speech interpretation, a set of 2800 utterances was recorded on the robotic platform covering each request type. Different speakers, including non-native ones (7 non-native among 16 French speakers) were involved in this recording task. In these rather difficult conditions, our speech recognition system reaches an accuracy of 84.71% which is quite good considering first that sentences were uttered in a different context than the audio data used for training acoustic and phonetic models and secondly that among the speakers, near half of them were non native. Around 5% of improvements were reached after a re-estimation step using the Baum-Welch algorithm applied to a subset of well recognized utterances (1049 utterances correctly recognized during the first evaluation): 89% of accuracy (at word level) was obtained on the remaining 1751 utterances.

Table 5 Different types of interaction requests

Greetings/Starting interaction with the Robot
"Bonjour Jido, c'est moi Paul" ("Hi Jido it's me Paul")
Basic or more advanced movement requests including deictic
"Tourne à gauche" ("Turn left")/"Viens ici" ("Come here")
Guidance request in the Human environment
"Emmène-moi à la salle robotique" ("Take me to the robotic room")
Interaction for object exchange
"Donne-moi cette bouteille" ("Give me this bottle")
Agreement/Disagreement/Thanks
"Oui" ("Yes")/"Non" ("No")/"Merci" ("Thank you")

In Austermann et al. (2010), an HRI application aimed at making a robot understand natural commands uttered by the robot's user, recognition is also based on the same speech recognition engine. Even if the underlying goal is different, mainly focusing on learning commands, the average recognition accuracy reached is between 80 and 84.5%.

5.1.2 Interpreting speech with RECO

In our experimental context, we focus on features like actions, objects, object attributes, location or robot configuration parameters. Extracting relevant information from each user utterances consists in interpreting the best speech recognition hypothesis. This is classically done by considering the word interpretations given by the semantic lexicon, which links words and their meaning in the scope of the application. These interpretations are then combined according to the syntactic word phrases described by the grammars, which in turn are combined to build the full sentence interpretation. For example, a sentence like "Pick up the red bottle" will be interpreted by combining the interpretation of the main two parts of such basic commands: the action to be performed and the object on which the action has to be performed. In our semantic lexicon "Pick up" will be describe by [action = GET_OBJECT; object = *unknown*], "red" by [color = RED], "bottle by [object = [name = BOTTLE]]. The interpretation of the phrase "the red bottle" will combine the interpretation of "red" and "bottle" to give [object = [name = BOTTLE; color = RED]] which in turn will instantiate the *unknown* value. The interpretation of this sentence will be [action = GET_OBJECT; object = [name = BOTTLE; color = RED]].

Then the result is compared with interpretation models. If one of them matches and is fully instantiated (each feature being assigned with a value), a valid and understandable command is generated and sent to the robot supervisor in order to be executed. Deictic and anaphoric words are defined in our semantic lexicon as related to a location which will be given by a gesture or the user position. This is specified by a semantic feature like [location = *FROM_GEST()*] or [ref_location = *GET_USER_POSITION()*]. So when a spatial reference is missing, fusion with gesture recognition results is required. Problems related to user intentions, dialog management and dialog strategies are not taken into account in this version. To evaluate this part of the processing, we computed the Correct Interpretation Rate (CIR) which reached 74.05 % of utterances correctly interpreted (resp. 81.8%) before (resp. after) acoustic model reestimation. Considering only the interpretation of the utterances requiring a complementary gesture, the CIR reaches 89.7%.

Once speech is recognized and interpreted by our embedded module, fusion can be performed with gesture recognition results.

5.2 Speech and gesture fusion

5.2.1 Related work

Natural communication performed by humans can be only verbal, when speech is sufficient in itself to convey a specific message like: "Pick up the red bottle", or can also be accompanied by a complementary gesture when saying for example: "Put this down here" associated with a pointing gesture. Even if some gestures can be significant enough regarding the context in which they are performed, sometimes uttering something will help to strengthen the gesture interpretation regarding the interaction context. For instance, when we naturally wave one hand, this means either "hello" or "goodbye" depending on the interaction situation. In these cases, a late fusion is operated, at a semantic level, once relevant pieces of information have been extracted by each component. Such a semantic fusion can be carried out in two steps as presented in Lopez-Cozar Delgado and Araki (2005): once events from each modality have been separately recognized and have been "literally" interpreted (interpretation at a low-level), they can be combined to perform a contextual interpretation (interpretation at a high-level) in order to extract the global meaning of speech and/or gesture. This semantic fusion process can also be applied to modalities characterized by different time scales. In this case, interpretation must be handled within a time window in which the temporal relationship between both modalities is considered as significant. The most advanced audio-visual fusion strategy in the HRI literature is described in Stiefelhagen et al. (2004): the two N-best recognition (speech and deictic mono-manual gestures) lists are merged via a hierarchical strategy. Our fusion process follows the same approach extended to symbolic gestures performed by moving one or both hands. Gesture recognition has been described in Sect. 4.3.

5.2.2 Merging semantic interpretation of speech and gesture with FUSION

As previously explained, using a hierarchical approach means that, if speech interpretation has detected the need of a complementary gesture, we proceed to a late fusion operated at the semantic level with the gesture interpretation results obtained within the same time window. Considering the speech interpretation I_{RECO} of the sentence uttered (*"Put the bottle down here"* and the interpretation I_{GEST} of the pointing gesture performed in the same time window, the fusion result will be obtained as follows

$$I_{RECO} = \begin{cases} \text{action} = \text{PUT} \\ \text{object} = [\text{name} = \text{BOTTLE}] \\ \text{location} = \text{FROM\_GEST}() \end{cases}$$

$$I_{GEST} = \{\text{location} = \text{POS}(X, Y, Z)$$

$$\text{FUSION}(I_{GEST}, I_{RECO}) = \begin{cases} \text{action} = \text{PUT} \\ \text{object} = [\text{name} = \text{BOTTLE}] \\ \text{location} = \text{POS}(X, Y, Z) \end{cases}$$

We first based the fusion step on the one-best hypothesis of each interpretation process, speech and mono or bimanual 3D gesture. But according to the interaction context, it often occurs that when the correct sentence or the correct gesture is not the best one, it can be found among the N-best hypotheses. To improve fusion results we propose, as in Stiefelhagen et al. (2004), to take N-best lists into account and compute a confidence score using the recognition score of each considered hypothesis.

5.2.3 Computing confidence scores for multimodal fusion robustness

Confidence scores can give precision about the reliability of information extracted from each modality. Lists of N-Best hypotheses can be easily produced by each recognition process. The number of hypotheses N_S for speech and N_G for gesture were respectively set to 10 and 12. These values were empirically chosen in order to take into account only the most useful information and limit the computation time. For speech for example, considering more than the ten first hypotheses will introduce some "noise" in the process, the words being lexically too far from the words uttered.

Considering h_i and g_j a speech and a gesture recognition hypothesis ($i \in [1, N_S]$ and $j \in [1, N_G]$). Each hypothesis comes along with its recognition score (log likelihood)

Table 6 Confusion matrix on fusion results (in %)

Gesture + speech utterance type	1	2	3	4	5	6	7	Others
1: "introducing oneself" + presentation	91	0	0	0	0	0	0	9
2: "greetings" (one or two hands) + hello-like	0	82	0	0	0	0	0	18
3: "stop" + stop-like	0	0	64	0	0	0	18	18
4: pointing gesture + "take this object"-like	0	0	0	91	0	0	0	9
5: "come to me" (one or two hands) + "come to me"-like	0	0	0	0	100	0	0	0
6: pointing gesture + "goto there"-like	0	0	0	0	0	100	0	0
7: "go away" + "go away"-like	0	0	9	0	0	0	91	0

$\log L(h_i)$ or $\log L(g_j)$. For each hypothesis, a confidence score is obtained as follows. A score $L(h_i)$ is computed from the normalized log likelihood of each speech hypothesis h_i

$$L(h_i) = \exp\left\{-\frac{(\frac{\log L(h_i)}{\overline{\log L}})^{N_S}}{\sigma_1}\right\}, \tag{8}$$

with $\overline{\log L} = \sum \frac{\log L(h_i)}{N_S}$, the mean of all the initial speech recognition scores, σ_1 a predefined standard deviation (here 0.2). For each word w appearing in the N-best hypotheses, a word confidence score $CS(w)$ is computed in order to strengthen the speech hypothesis. The more often the word appears in speech hypotheses, higher the score is. This score is given by

$$CS(w) = \frac{\sum_{w \in h_i} L(h_i)}{\sum_w L(h_i)}. \tag{9}$$

The final confidence score associated with each speech hypothesis h_i and given by (10) becomes

$$S(h_i) = \frac{L(h_i)}{NW_i} \cdot \frac{\sum\limits_{w \in h_i} CS(w)}{\sum\limits_w CS(w)}, \tag{10}$$

with NW_i, the number of words in the ith hypothesis.

Confidence scores on N-best gesture recognition results are obtained as follows. A score $L(g_j)$ is computed from the normalized log likelihood of each gesture hypothesis g_j:

$$L(g_i) = \exp\left\{-\frac{(\frac{\log L(g_j)}{\overline{\log L}})^{N_G}}{\sigma_2}\right\}, \tag{11}$$

with $\overline{\log L} = \sum \frac{\log L(g_j)}{N_G}$, the mean of all gesture recognition scores and σ_2 a predefined standard deviation (here 0.5).

Finally a fusion score is computed for each speech recognition hypothesis h_i, to take into account a complementary gesture: if there is a gesture g_j in the same time window that can complement h_i, a new score L_f is associated to h_i using (12) else (13) is used.

$$L_f(h_i) = L(g_j)^{\alpha} \cdot S(h_i)^{(1-\alpha)}, \tag{12}$$

$$L_f(h_i) = {L_M}^{\alpha} \cdot S(h_i)^{(1-\alpha)} \tag{13}$$

where $\alpha = \frac{1}{2}\frac{T_g}{T_s}$ is used to take into account the difference of recognition rate (T_g and T_s respectively for gesture and speech, obtained in off-line experiments) between the two modalities and $L_M = \exp\{-\frac{1}{\sigma_2}\}$.

In (13), that is when no gesture is needed, we act as if there were a gesture whose score is equal to $\overline{\log L}$.

To compare these new results with those coming from each independent interpretation obtained before any fusion process, we compute a confusion matrix presented in Table 6. Each type of gesture among the seven categories identified is played ten times along with the corresponding speech request. Among theses 70 multimodal requests performed, two are misinterpreted and mixed up with another multimodal request category, while 5 of them are not interpreted as multimodal but as monomodal (column "others" stands for "only speech been detected"). Confusion is also very important between types 3 and 7, which can easily be explained. On one hand, the two corresponding gestures are very similar, and on the other hand, each "Stop"-like request ("stop" and other corresponding formulations in French) being quite short, speech recognition does not perform very well when they are uttered alone. In other experiments, not described here, specific acoustic and phonetic models have been built for each critical French short words (like "stop", "yes", "no", ...) which has improved their recognition, but this was not used in the experiments described here. Nevertheless, the fusion performed as presented in this section clearly improves the interpretation of multimodal requests as the associated rate of correct interpretation reaches 92%. Recall that the correct speech interpretation alone was about 88% and the correct gesture interpretation about 71%. Other evaluations (not detailed here) have been performed in order to prove the robustness of our system with multiple users. The experiments involved four

different users performing 8 × 6 commands among which one was based on speech only. Every multimodal command succeeded, the only case of failure happened twice on the monomodal command, see Vallée et al. (2009) for more details. To validate our multimodal interface, live experiments have been set up on the robotic platform called Jido, previously described.

6 Robotic scenarios and associated live experiments

6.1 Scenario description

Our "human perception" modules encapsulated in the multimodal interface have been tested within multiple scenarios on our robot Jido. The goal is to raise interest and prove the efficiency of our modules for the interaction between a human user and a robot. These scenarios involve different kind of commands (speech only, speech with spatial references, speech + symbolic gesture, speech + deictic gesture). Also, the robot operates in diverse H-R situations (standing or seated human, human and robot moving in the room, etc.) and environments (diverse tables of different heights).

It is important to note that, if all these scenarios deal with taking/exchanging an object as well as more or less complex movements of the robot, our contribution (and what is of interest here) focuses on the ability of the robot to understand the user by means of our interface. Actually, the robot could not know which object to take, where to put it or where to go, if pointing gestures were not taken into account and combined with speech. It is also by combining vision and speech that object exchange with the user could be handled by the robot. Finally, it is through the use of symbolic gestures, in addition to speech, that the interaction can become both more natural (we often accompany our words with gestures, even unintentionally) and safe (their combination enables to correct errors/uncertainties of isolated channels). Another important point is that these experiments always stay in the framework of peer to peer interaction: the robot is not meant to deal with multiple people in our scenario. However, the system is quite capable of managing any interruption provoked by another person, the tracker being previously locked on the intended user, of course if this person does not mask the robot's sight of the user's hands.

For compactness reasons, the paper focuses on the scenario which is perhaps the most challenging: interaction tasks inducing large movements of both human and robot, tracker re-initialization, mono and bi-manual symbolic and deictic gestures, face identification (Table 7). The two other scenarios illustrated in Fig. 1 were otherwise tested and prove that our approach is robust and generic enough to work in diverse conditions. The interested reader will find videos of these scenarios/results at http://brice.burger.pagesperso-orange.fr/these/index.html.

Since in our scenarios we have to deal with misunderstanding on the robot side, we refer to the human-human communication and the way to cope with understanding failure. Being faced with such situations, a person generally resumes his/her latest request in order to be understood. In our scenario, although no real dialog management was implemented yet, we wanted to give the robot the possibility to ask the user to repeat his/her request each time one of the planed step fails without irreversible consequences. By saying: "I did not understand, please try again." (*via* the speech synthesis module named `Speak` see Fig. 3), the robot resumes the most recent step from the beginning. The multimodal interface runs completely onboard.

6.2 Experiments and results

From this key scenario, several experiments were conducted in our institute environment. The user asked Jido to follow his/her instructions given by means of multimodal requests by first asking Jido to come close to a given table, take over the pointed object and give it to him/her. Figure 5 illustrates the scenario execution. For each step, the main picture depicts the current H/R situation, while the subfigure shows the tracking results of the `GEST` module. In this trial, the multimodal interface succeeds to interpret multimodal commands and to safely manage object exchanges with the user. The entire video is joined with the paper, but is also available in full quality and with more illustrations at http://brice.burger.pagesperso-orange.fr/these/index.html.

Given this scenario, quantitative performance evaluations were also conducted. They refer to both (i) robot capacity to execute the scenario, (ii) potential user acceptance of the ongoing interaction scenario. The less failures occur, the more comfortable the interaction will be for the user. The associated statistics are summarized in Table 8 which synthesizes the data collected after 14 scenario executions.

Let us comment on these results. In 14 trials of the full scenario execution, we observe only 1 fatal failure (noted fatal) due to a motion planning error independently from our multimodal interface. Besides, we consider that a scenario run involving more than 3 failures is potentially unacceptable by the user who can be easily bored by being constantly asked to re-perform his/her request. These situations were encountered when pushing the limits of our system, for example when the precision of pointing gestures decreases with the angle between the head-hand line and the table. In the same manner, short utterances are still difficult to recognize especially when the environment is polluted with short sudden noises.

This system was designed in order to address untrained people. Even if the system is currently not fast and open

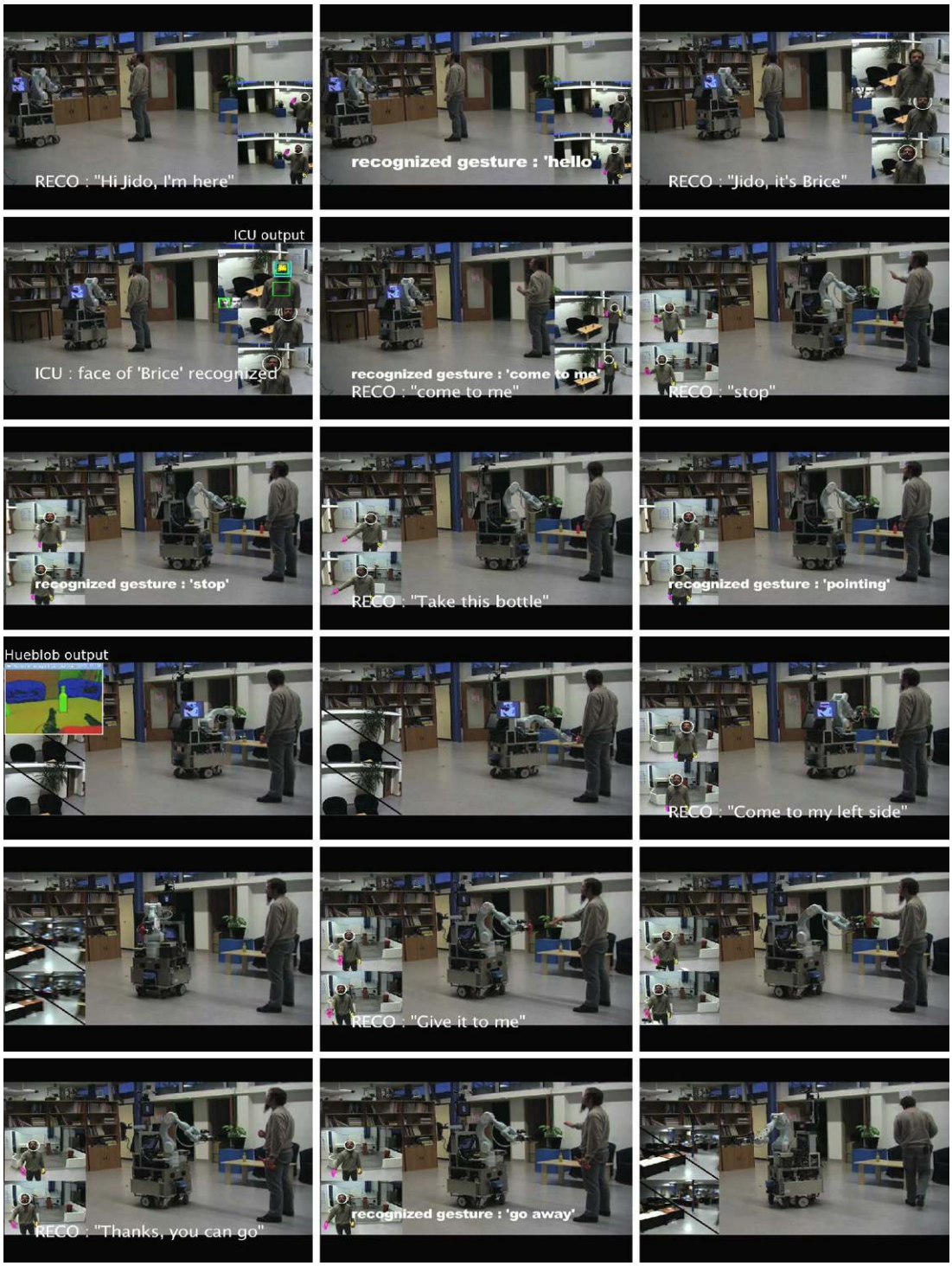

Fig. 5 Snapshots of a H/R scenario based on speech and gesture data fusion. From top left to bottom right, the user is calling the robot, then after introducing himself and being recognized by the robot (face recognition) he is asking the robot to come to him and take the object he is pointing at (object recognition). The robot is asked by the user to come to this left side to give him the object. Each frame shows the general view of the interaction situation as well as the GEST module results (two small images on the right or left corner). When other modules are used, their results are shown in a third image above the GEST results (face recognition in the third and fourth images and object recognition in the tenth image)

Table 7 Excerpt of an interaction scenario between a human user and Jido

#	Human user command	Jido action	Demonstrated modules	Comments
1.	*"Hello, I am here"* accompanied with a symbolic gesture	local motion towards the user	`RECO, GEST, FUSION, Motion planning modules`	Jido is disrupting its current task, moves and stops in front of him/her
2.	*"Hi Jido it's me Brice"*	face recognition	`RECO, Face recognition module`	The user must be beforehand identified to be allowed to interact with Jido
3.	*"Come to me"* with a symbolic gesture	local motion according to the user location	`GEST, RECO, FUSION, Motion planning modules`	The command execution requires the 3D location of the user
4.	*"Stop"* with a symbolic gesture	stop of the robot	`GEST, RECO, FUSION`	This command is performed while the robot is moving
5.	*"Take this object"* with a pointing gesture	grasping of the pointed object	`GEST, RECO, FUSION, Motion planning and Object recognition modules`	The robot searches for an object the user points to, then picks it up if present
6.	*"Go to my left"*	local motion according to the user location	`GEST, RECO, FUSION, Motion planning modules`	The command execution requires the 3D location of the user
7.	*"Give the object to me"*	giving the object in the user's hand	`GEST, RECO, FUSION, Motion planning modules`	The command execution requires no gesture recognition but "only" the hand tracking in order to put the object in the user's hand
8.	*"Go away"* with a symbolic gesture	local motion to go away from the user	`GEST, RECO, FUSION, Motion planning modules`	

Table 8 Modules failure rates during scenario trials

#	"RECO"	"GEST"	"FUSION"	Others	Comments
1.	0	1	0	0	
2.	0	0	0	1 Face recognition	face recognition
3.	1	3	1	0	the distance to the robot makes this gesture hard to track
4.	3	2	2	0	computing time sometimes too long when the robot is moving
5.	0	0	0	2 Object detection	the bottle is not always seen
6.	0	0	0	0	the left is not always really on the left...
7.	0	0	0	2 Motion planning (1 fatal)	hand too far, localization failure
8.	2	4	1	0	

enough for a use with real native users in its current state (see the video), the remaining step is only a matter of optimization. Actually, as described in Sect. 5.1, our pho-nemes were not trained on data recorded in our experimental context but just re-estimated using a small set of speech utterances recorded through the robotic platform These speech

sequences were uttered for half by non native French speakers. Speakers were not aware of how the system works and did not receive guidelines on what to say and how. We also based our grammar on natural speech, as varied as possible, in order to enable a relatively large freedom in the commands given to the robot.

On the other hand, as described in Sect. 4.3, our gestures have also been learned without strong constraints on their shape or speed, which allowed people to perform them in a natural way. This diversity in the learning data set creates a real freedom in how to make a gesture in front of the robot, the downside being a greater similarity between gestures and thus a lower recognition rate than with highly codified gestures.

Apart from these limitations, the multimodal interface is shown to be robust enough to allow continuous operations for the long term experimentations that are intended to be performed.

7 Conclusion

This article described a multimodal interface for natural interaction between human users and a mobile robot. The first contribution concerns a bank of distributed particle filters for the simultaneous tracking of two-handed gestures and head tracking in 3D. Our IIDMOT is claimed to solve data association and "coalescence" problems, as well as automatic filter (re)-initialization, when the targeted limbs are in close proximity, become occluded or transiently exit the camera field of view. These situations usually occur when performing natural gestures with or without the dominant hand (or with both hands) in front of a mobile platform. Finally, the strategy is shown to be person independent, less time consuming (compared to Human-Machine Communication systems) while our hybrid data fusion principle (based on both appearance and 3D cues) is shown to improve the tracker versatility and robustness to clutter.

The second contribution concerns gesture and speech probabilistic fusion at the semantic level. We use an open source speech recognition engine (Julius) for speaker independent recognition of continuous speech. Speech interpretation is done on the basis of the N-best speech recognition results and a confidence score is associated with each hypothesis. By this way, we strengthen the reliability of our speech recognition and interpretation processes. Results on pre-recorded data illustrated the high level of robustness and usability of our interface. Clearly, it is worthwhile to augment the gesture recognizer with a speech-based interface as the robustness reached by cue proper fusion is much higher than for single cues.

Finally, the third contribution concerns robotic experiments which illustrated a high level of robustness and usability of our interface. While this is only a key scenario designed to test our interface, we think that the latter opens an increasing number of interaction possibilities. To our knowledge, few mature robotic systems benefit from such advanced embedded multimodal interaction capabilities.

Several directions are currently studied regarding this multimodal interface. First, our tracking modality will be made much more active. Zooming will be used to actively adapt the focal length with respect to the H/R distance and the current robot status. A second envisaged extension is, in the vein of Richarz et al. (2006), Stiefelhagen et al. (2004), to incorporate the head orientation as additional features in the gesture characterization as it is a common fact that people generally look at the target they are pointing to. The gesture recognition performances and the precision of the pointing direction should be increased significantly. Further investigations will aim to augment the gesture vocabulary and refine the fusion process, between speech and gesture. The major computational bottleneck will become the gesture recognition process.

Acknowledgements The work described in this paper was partially conducted within the EU Project CommRob ("Advanced Robot behavior and high-level multimodal communication"—www.commrob.eu) under contract FP6-IST-045441 and the French ANR project AMORCES.

References

Alami, R., Chatila, R., Fleury, S., & Ingrand, F. (1998). An architecture for autonomy. *The International Journal of Robotics Research*, *17*(4), 315–337.

Arras, K., & Burgard, W. (Eds.), *Robots in exhibitions*, Lausanne, Switzerland, October 2002.

Austermann, A., Yamada, S., Funakoshi, K., & Nakano, M. (2010). Learning naturally spoken commands for a robot. In *Interspeech*, Makuhari, Japan, September 2010.

Axenbeck, T., Bennewitz, M., Behnke, S., & Burgard, W. (2008). Recognizing complex, parameterized gestures from monocular image sequences. In *IEEE-RAS international conference on humanoid robots (Humanoids'08)*, Daejeon, South Korea, December 2008.

Azad, P., Ude, A., Asfour, T., & Dillman, R. (2007). Stereo-based markerless human motion capture for humanoid robot systems. In *Int. conf. on robotics and automation (ICRA'07)*, Roma, Italy, April 2007.

Badii, A., & Thiemert, D. (2009). The CompanionAble project. In *Workshop co-located with the Europ. conf. on ambient intelligence*, Salzburg, Austria, November 2009.

Bar-Shalom, Y., & Jaffer, A. G. (1998). *Tracking and data association*. San Diego: Academic Press.

Bennewitz, M., Faber, F., Joho, D., Schreiber, M., & Behnke, S. (2005). Towards a humanoid museum guide robot that interacts with multiple persons. In *Int. conf. on humanoid robots (HUMANOID'05)* (pp. 418–423). Tsukuba, Japan.

Bernier, O., & Collobert, D. (2001). Head and hands 3D tracking in real-time by the EM algorithm. In *Workshop of int. conf. on computer vision*, Vancouver, Canada.

Bischoff, R., & Graefe, V. (2004). HERMES—a versatile personal robotic assistant. *Proceedings of the IEEE*, *92*, 1759–1779.

Chen, F. S., Fu, C. M., & Huang, C. L. (2003). Hand gesture recognition using a real-time tracking method and hidden Markov models. *Image and Vision Computing*, *21*(8), 745–758.

Corradini, A., & Gross, H. M. (2000). Camera-based gesture recognition for robot control. In *Int. joint conf. on neural networks (IJCNN'00)*, Roma, Italy, July 2000.

Davis, F. (1971). *Inside intuition-what we know about non-verbal communication*. New York: McGraw-Hill.

Erol, A., Bebis, G., Nicolescu, M., Boyle, R., & Twombly, X. (2007). Vision-based hand pose estimation: a review. *Computer Vision and Image Understanding*, *108*, 52–73.

Fels, S., & Hinton, G. (1997). Glove-talk II: A neural network interface which maps gestures to parallel format speech synthesizer controls. *IEEE Transactions on Neural Networks*, *9*(1), 205–212.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, *42*, 143–166.

Fontmarty, M., Lerasle, F., & Danès, P. (2007). Data fusion within a modified annealed particle filter dedicated to human motion capture. In *Int. conf. on intelligent robots and systems (IROS'07)* (pp. 3391–3396). San Diego, USA, November 2007.

Fox, M., Ghallab, M., Infantes, G., & Long, D. (2006). Robot introspection through learned hidden Markov models. *Artificial Intelligence*, *170*(2), 59–113.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J. F., & Gravier, G. (2005). The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *Interspeech/Eurospeech*, Lisbon, Portugal, September 2005.

Gorostiza, J., Barber, R., Khamis, A., & Malfaz, M. (2006). Multimodal human-robot interaction framework for a personal robot. In *Int. symp. on robot and human interactive communication (RO-MAN'06)* (pp. 39–44). Hatfield, UK, September 2006.

Hanafiah, Z. M., Yamazaki, C., Nakamura, A., & Kuno, Y. (2004). Human-robot speech interface understanding inexplicit utterances using vision. In *CHI 2004* (pp. 1321–1324). Vienna, Austria, April 2004.

Harte, E., & Jarvis, R. (2007). Multimodal human-robot interaction in an assistive technology context. In *Australian conf. on robotics and automation*, Brisbane, Australia, December 2007.

Hasanuzzaman, M., Ampornaramveth, V., Zhang, T., Bhuiyan, M., Shirai, Y., & Ueno, H. (2004). Real-time vision-based gesture recognition for human robot interaction. In *Int. conf. on robotics and biomimetics*, Shenyang, China, August 2004.

Hasanuzzaman, M., Zhang, T., Ampornaramveth, V., & Ueno, H. (2007). Adaptive visual gesture recognition using a knowledge-based software platform. *Robotics and Autonomous Systems*, *55*(8), 643–657.

Huang, Y., Huang, T., & Niemann, H. (2002). Two-handed gesture tracking incorporating template warping with static segmentation. In *Int. conf. on automatic face and gesture recognition (FGR'02)*, Washington, USA, May 2002 (pp. 275–280).

Isard, M., & Blake, A. (1998a). I-CONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. In *European conf. on computer vision (ECCV'98)* (pp. 893–908). Freibourg, Germany, June 1998.

Isard, M., & Blake, A. (1998b). CONDENSATION—conditional density propagation for visual tracking. *International Journal of Computer Vision*, *29*(1), 5–28.

Isard, M., & Blake, A. (2001). BraMBLe: a Bayesian multiple blob tracker. In *Int. conf. on computer vision (ICCV'01)* (pp. 34–41). Vancouver, Canada.

Just, A., Marcel, S., & Bernier, O. (2004). HMM and IOHMM for the recognition of mono and bi-manual 3D hand gestures. In *British machine vision conference (BMVC'04)*, London, UK, September 2004.

Lee, A., Kawahara, T., & Shikano, K. (2001). Julius—an open source real-time large vocabulary recognition engine. In *European conference on speech communication and technology (EUROSPEECH)* (pp. 1691–1694). Aalborg, Denmark, September 2001.

Lopez-Cozar Delgado, R., & Araki, M. (2005). *Spoken, multilingual and multimodal dialogues systems—development ans assessment*. New York: Wiley.

Maas, J. F., Spexard, T., Fritsch, J., Wrede, B., & Sagerer, G. (2006). BIRON, what's the topic? a multi-modal topic tracker for improved human-robot interaction. In *Int. symp. on robot and human interactive communication (RO-MAN'06)*, Hatfield, UK, September 2006.

Moeslund, T., Hilton, A., & Kruger, V. (2006). A survey of advanced vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, *104*, 174–192.

Murphy-Chutorian, E., & Trivedi, M. (2008). Head pose estimation in computer vision: a survey. *Transactions on Pattern Analysis Machine Intelligence (PAMI'08)*.

Nickel, K., & Stiefelhagen, R. (2006). Visual recognition of pointing gestures for human-robot interaction. *Image and Vision Computing*, *3*(12), 1875–1884.

Park, H. S., Kim, E. Y., Jang, S., & Park, S. H. (2005). HMM-based gesture recognition for robot control. In *Iberian conf. on pattern recognition and image analysis (IbPRIA'05)*, Estoril, Portugal, June 2005.

Pérennou, G., & de Calmès, M. (2000). MHATLex: Lexical resources for modelling the French pronunciation. In *Int. conf. on language resources and evaluations* (pp. 257–264). Athens, Greece, June 2000.

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

Pineau, J., Montemerlo, M., Pollack, M., Roy, N., & Thrun, S. (2003). Towards robotic assistants in nursing homes: challenges and results. *Robotics and Autonomous Systems*, *42*, 271–281.

Prodanov, P., & Drygajlo, A. (2003a). Multimodal interaction management for tour-guide robots using Bayesian networks. In *Int. conf. on intelligent robots and systems (IROS'03)* (pp. 3447–3452). Las Vegas, Canada, October 2003.

Prodanov, P., & Drygajlo, A. (2003b). Bayesian networks for spoken dialogue managements in multimodal systems of tour-guide robots. In *European conf. on speech communication and technology (EUROSPEECH'03)* (pp. 1057–1060). Geneva, Switzerland. September 2003.

Qu, W., Schonfeld, D., & Mohamed, M. (2007). Distributed Bayesian multiple-target tracking in crowded environments using multiple collaborative cameras. *EURASIP Journal on Advances in Signal Processing*.

Rasmussen, C., & Hager, G. (2001). Probabilistic data association methods for tracking complex visual objects. *Transactions on Pattern Analysis Machine Intelligence* 560–576.

Richarz, J., Martin, C., Scheidig, A., & Gross, H. M. (2006). There you go!—estimating pointing gestures in monocular images for mobile robot instruction. In *Int. symp. on robot and human interactive communication (RO-MAN'06)* (pp. 546–551). Hartfield, UK, September 2006.

Rogalla, O., Ehrenmann, M., Zollner, R., Becher, R., & Dillman, R. (2004). Using gesture and speech control for commanding a robot. In *Advances in human-robot interaction* (Vol. 14). Berlin: Springer.

Shimizu, M., Yoshizuka, T., & Miyamoto, H. (2006). A gesture recognition system using stereo vision and arm model fitting. In *Int. conf. on brain-inspired information technology (BrainIT'06)*, Hibikino, Japan, September 2006.

Siegwart, R. et al. (2003). Robox at expo 0.2: a large scale installation of personal robots. *Robotics and Autonomous Systems*, *42*, 203–222.

Skubic, M., Perzanowski, D., Blisard, S., Schultz, A., & Adams, W. (2004). Spatial language for human-robot dialogs. *IEEE Transactions on Systems, Man, and Cybernetics*, *2*(34), 154–167.

Stiefelhagen, R., Fügen, C., Gieselmann, P., Holzapfel, H., Nickel, K., & Waibel, A. (2004). Natural human-robot interaction using speech head pose and gestures. In *Int. conf. on intelligent robots and systems (IROS'04)*, Sendal, Japan, October 2004.

Stückler, J., Gräve, K., Kläß, J., Muszynski, S., Schreiber, M., Tischler, O., Waldukat, R., & Behnke, S. (2009). Dynamaid: Towards a personal robot that helps with household chores. In *Robotics: science and systems conference (RSS'09)*, Seattle, USA, June 2009.

Thayananthan, A., Stenger, B., Torr, P. H. S., & Cipolla, R. (2003). Learning a kinematic prior for tree-based filtering. In *British machine vision conf. (BMVC'03)* (Vol. 2, pp. 589–598). Norwick, UK, September 2003.

Theobalt, C., Bos, J., Chapman, T., & Espinosa, A. (2002). Talking to godot: Dialogue with a mobile robot. In *Int. conf. on intelligent robots and systems (IROS'02)*, Lausanne, Switzerland, September 2002.

Triesch, J., & Von der Malsburg, C. (2001). A system for person-independent hand posture recognition against complex backgrounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(12), 1449–1453.

Vallée, M., Burger, B., Ertl, D., Lerasle, F., & Falb, J. (2009). Improving user of interfaces robots with multimodality. In *Int. conf. on advanced robotics (ICAR'09)*, Munich, Germany.

Viola, P., & Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. In *Int. conf. on computer vision and pattern recognition (CVPR'01)*, Hawaii, December 2001.

Waldherr, S., Thrun, S., & Romero, R. (2000). A gesture-based interface for human-robot interaction. *Autonomous Robots*, *9*(2), 151–173.

Yoshizaki, M., Kuno, Y., & Nakamura, A. (2002). Mutual assistance between speech and vision for human-robot interface. In *Int. conf. on intelligent robots and systems (IROS'02)* (pp. 1308–1313). Lausanne, Switzerland, September 2002.

Yu, T., & Wu, Y. (2004). Collaborative tracking of multiple targets. In *Int. conf. on computer vision and pattern recognition (CVPR'04)*, Washington, USA, June 2004.

Zhao, T., & Nevatia, R. (2004). Tracking multiple humans in crowded environment. In *Int. conf. on computer vision and pattern recognition (CVPR'04)*, Washington, USA, June 2004.

B. Burger received an engineering diploma in Industrial Risks Management from ENSIB (Bourges, France) and a master's degree in Sensors, Electronic and Robotic Systems from the University of Versailles-Saint-Quentin (France) in 2006. He defended his Ph.D. in 2010 at the University of Toulouse (France).
His work, carried out at LAAS-CNRS and IRIT labs (Toulouse, France), was mainly about extracting audio and visual data from the robot sensors, and fuse them in order to enable advanced human-robot interaction.

I. Ferrané defended her Ph.D. in computer science in 1991 at Paul Sabatier University where she has been an assistant professor since 1993. Her thesis carried out at IRIT (Computer science Institute of Toulouse) was about lexical and morphosyntactic databases for spoken language.
Her research activities in Samova team at IRIT focus now on high level event detection from multimedia content. By exploiting and combining low or mid level features this work aims at extracting, characterising and even interpreting high level events. Two main fields are concerned: (1) audiovisual content indexing and structuring by studying interaction between speakers, searching for speaker roles, interaction sequences like interview or debate in order to focus on conversational speech as well as (2) multimodal human-robot interaction using speech and gesture modalities.

F. Lerasle is an assistant professor at Paul Sabatier Universitysince September 1997, and researcher at LAAS-CNRS in vision forrobotics in Toulouse. His PhD thesis was on human motion capture by multiocular vision at the LASMEA, graduating from Blaise Pascal University of Clermont-Ferrand in 1997.
His current research at LAAS-CNRS concerns vision for robotics, more particularly: (1) detection, recognition, tracking of people, as well as interpretation of their gestures and activities for human-robot interaction, (2) landmark detection/recognition for metrical ortopological navigation of mobile robots in indoor environments. He is the author or co-author of a fifty scientific papers in international conferences or journals, most of them in the field of human perception from mobile robotics.

G. Infantes received a master's degree in computer science and applied mathematics from ENSEEIHT, Toulouse, France in 2002 and two master's degree from the University of Toulouse, France in 2002 and 2003. He received a Ph.D. from the University of Toulouse in computer systems in 2006, and worked at LAAS/CNRS, one of the largest computer science laboratories in France.
His work was mainly about modelling and controlling automous behaviors for robotic systems. He spent a year as a research assistant at the University of Maryland's Institute for Advanced Computer Studies (UMIACS). G. Infantes has worked at ONERA (France) in autonomous systems area since 2008.