Multimodal Interaction Abilities for a Robot Companion

Brice Burger[1,2,3], Isabelle Ferrané[2,3], and Frédéric Lerasle[1,3]

[1] CNRS ; LAAS ; 7, avenue du Colonel Roche, F-31077 Toulouse, France
[2] IRIT ; 118 route de Narbonne, F-31077 Toulouse, France
[3] Université de Toulouse ; UPS, INSA, INP, ISAE ; LAAS-CNRS : F-31077 Toulouse, France
{burger,ferrane}@irit.fr, lerasle@laas.fr

Abstract. Among the cognitive abilities a robot companion must be endowed with, human perception and speech understanding are both fundamental in the context of multimodal human-robot interaction. In order to provide a mobile robot with the visual perception of its user and means to handle verbal and multimodal communication, we have developed and integrated two components. In this paper we will focus on an interactively distributed multiple object tracker dedicated to two-handed gestures and head location in 3D. Its relevance is highlighted by in- and off- line evaluations from data acquired by the robot. Implementation and preliminary experiments on a household robot companion, including speech recognition and understanding as well as basic fusion with gesture, are then demonstrated. The latter illustrate how vision can assist speech by specifying location references, object/person IDs in verbal statements in order to interpret natural deictic commands given by human beings. Extensions of our work are finally discussed.

Keywords: particle filtering, multiple object tracking, speech understanding, multimodal interaction, personal robotics.

1 Introduction and Framework

The development of socially interactive robots is a motivating challenge, so that a considerable number of mature robotic systems have been developed during the last decade [3]. Moving such robots out of laboratories, *i.e.* in private homes, to become robot companions is a deeper challenge because robots must be endowed with cognitive abilities to perform a unconstrained and natural interaction with non-expert users. Besides the verbal information, gestures and reactive body motions stemmed from audio and video stream analysis must also be considered to achieve a successful intuitive communication/interaction with a household autonomous platform. This also raises issues related to efficiency and versatility. Because of the concurrent execution of other embedded functions, only a small percentage of the robot's computational power can be allocated to the interactive system. Meanwhile, as the on-board sensors are moving instead of being static, the interactive system is faced with noisy and cluttered environments.

A. Gasteratos, M. Vincze, and J.K. Tsotsos (Eds.): ICVS 2008, LNCS 5008, pp. 549–558, 2008.

On one hand, fusing the interpretation of auditive and visual features improves the system robustness to such environments. On the other hand, their combination, permits to specify parameters related to person/object IDs or location references in verbal statements, typically *"give him a glass", "give this object to me", "put it there"*. Many interactive robotic systems make use of single-hand gesture [7,8,10] and/or object recognition [7,10] to complete the message conveyed by the verbal communication channel. Considering at once the identification of a person's face and pose, as well as two-handed gestures, must clearly disambiguates verbal utterances and so enriches any H/R interaction mechanism. An essential issue that we want to address in this context concerns the design of body and gesture trackers which must be endowed with both properties: visual data fusion (in the vein of [8]) and automatic re-initialization. All this makes our trackers work under a wide range of viewing conditions and aid recovery from transient tracking failure, which are due for instance to out-field of sight when the user is performing gestures.

The paper is organized as follows. Section 2 presents our particle filtering framework for the binocular tracking of multiple targets, namely the user's head and two-handed gestures. Section 3 presents preliminary robotic experiments involving involving this component and the one that is in charge of verbal and multimodal communication. Last, section 4 summarizes our contributions and discuss future extensions.

2 Visual Perception of the Robot User

2.1 3D Tracking of Heads and Hands

Our system dedicated to the visual perception of the robot user includes 3D face and two-hand tracking. Particle filters (PF) constitute one of the most powerful framework for view-based multi-tracking purpose [12]. In the robotics context, their popularity stems from their simplicity, modeling flexibility, and ease of fusion of diverse kinds of measurements. Two main classes of multiple object tracking (MOT) can be considered. While the former, widely accepted in the Vision community, exploits a single joint state representation which concatenates all of the targets' states together [6], the latter uses distributed filters, namely one filter per target. The main drawback of the centralized approach remains the number of required particles which increases exponentially with the state-space dimensionality. The distributed approach, which is the one we have chosen, suffers from the well-known "error merge" and "labeling" problems when targets undergo partial or complete occlusion. In the vein of [12], we develop a interactively distributed MOT (IDMOT) framework which 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 $z_{1:t}$. A linear point-mass combination

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

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

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 Euclidian distance exceeds a predefined threshold (annoted d_{TH} 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 [12], the repulsion "weight" $\varphi_1(.)$ follows

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

with β_1 and σ_1 two normalization terms being determined *a priori*. $D_{i,n}$ terms the Euclidian distance between particle $\mathbf{x}_t^{i,n}$ and temporary particle $\mathbf{x}_{t,k}^j$. The principle can be extended to 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

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

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 follows this principle but is extended in three ways. First, the conventional CONDENSATION [4] strategy is replaced by the ICONDENSATION [5] one whose importance function $q(.)$ in step 3 of Table 1 permits automatic (re)-initialization when the targeted human body parts appear or reappear in the scene. The principle consists in sampling the particle according to visual detectors $\pi(.)$, dynamics $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, and the prior p_0 so that, with $\alpha \in [0; 1]$

$$q(\mathbf{x}_t^{i,n}|\mathbf{x}_{t-1}^{i,n}, z_t^i) = \alpha\pi(\mathbf{x}_t^{i,n}|z_t^i) + (1-\alpha)p(\mathbf{x}_t^i|\mathbf{x}_{t-1}^{i,n}). \tag{3}$$

Secondly, the IDMOT particle filter, devoted initially to the image-based tracking of multiple objects or people, 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 - enables to benefit both from reconstruction-based and appearance-based approaches. The aim of our IDMOT approach, named IIDMOT, is to fit the projections all along the video stream of a sphere and two deformable ellipsoids (resp. for the head and the two hands), through the estimation of the 3D location $\mathcal{X} = (X, Y, Z)^{'}$, the orientation $\Theta = (\theta_x, \theta_y, \theta_z)^{'}$, and the axis length [1] $\Sigma = (\sigma_x, \sigma_y, \sigma_z)^{'}$ 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 $p(\mathbf{x}_l^i|\mathbf{x}_{t-1}^i)$, 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

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

Table 1. Our IIDMOT algorithm

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**
IF $t \geq 1$ **THEN** $\{$—$[\{\mathbf{x}_{t-1}^{i,n}, w_{t-1}^{i,n}\}]_{n=1}^N$ being a particle description of $p(\mathbf{x}_{t-1}^i | z_{1:t-1}^i)$—$\}$
"Propagate" the particle $\{\mathbf{x}_{t-1}^{i,n}\}_{n=1}^N$ by independently sampling $\mathbf{x}_t^{i,n} \sim q(\mathbf{x}_t^i | x_{t-1}^{i,n}, z_t^i)$
Update the weight $\{w_t^{i,n}\}_{n=1}^N$ associated to $\{\mathbf{x}_t^{i,n}\}_{n=1}^N$ according to the formula $w_t^{i,n} \propto w_{t-1}^{i,n} \frac{p(z_t^i | \mathbf{x}_t^{i,n}) p(\mathbf{x}_t^{i,n} | \mathbf{x}_{t-1}^{i,n})}{q(\mathbf{x}_t^{i,n} | \mathbf{x}_{t-1}^{i,n}, z_t^i)}$, prior to a normalization step so that $\sum_n w_t^{i,n} = 1$
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 $p(\mathbf{x}_t^i | z_{1:t}^i)$
FOR $j = 1 : i$, **DO**
  IF $d_{ij}(\hat{\mathbf{x}}_{t,k}^i, \hat{\mathbf{x}}_{t,k}^j) < d_{TH}$ **THEN**
    Save link(i,j)
    FOR k=1:K iterations, **DO**
      Compute φ_1, φ_2
      Reweight $w_t^{i,n} = w_t^{i,n}.\varphi_1.\varphi_2$
      Normalization step for $\{w_t^{i,n}\}_{n=1}^N$
      Compute the MMSE estimate $\hat{\mathbf{x}}_t^i$
      Compute φ_1, φ_2
      Reweight $w_t^{j,n} = w_t^{j,n}.\varphi_1.\varphi_2$
      Normalization step for $\{w_t^{j,n}\}_{n=1}^N$
      Compute the MMSE estimate $\hat{\mathbf{x}}_t^j$
    END FOR
  END IF
END FOR
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 $\frac{1}{N}$
END IF

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 $\pi(.)$ (equation (3)). These are also drawn from Gaussian distribution for head or hand configuration but deduced from skin color blob segmentation in the stereo video stream. The centroids and associated covariances of the matched regions are finally triangulated using the parameters of the calibrated stereo setup. For the weight updating step, each ellipsoid defined by its configuration $\mathbf{x}_t^i$ is then projected in one of the two image planes. Given $Q = \begin{bmatrix} A & b \\ b' & c \end{bmatrix}$ 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. For each ellipsoid projection, the pixels in the image are partitioned into a set of target pixels O, and a set of background pixels B. Assuming pixel-wise independence, the skin color-based likelihood is factored as

$$p(z_t^{i,c} | \mathbf{x}_t^i) = \prod_{o \in O} p_s(o | \mathbf{x}_t^i) \prod_{b \in B} [1 - p_s(b | \mathbf{x}_t^i)], \tag{4}$$

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 be moving, at

least intermittently. We thus favor the moving edges (if any) of the target in this likelihood so that

$$p(z_t^{i,s}|\mathbf{x}_t^i) \propto \exp\left(-D^2/2\sigma_s^2\right),\ D = \sum_{j=1}^{N_p} |x(j) - z(j)| + \rho\gamma(z(j)), \tag{5}$$

which 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

$$p(z_t^{i,c}, z_t^{i,s}|\mathbf{x}_t^i) = p(z_t^{i,c}|\mathbf{x}_t^i).p(z_t^{i,s}|\mathbf{x}_t^i). \tag{6}$$

2.2 Experimental Results

Prior to their integration on our mobile robot, experiments on a database 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, (ii) characterize its performances. This sequence set involves variable viewing conditions, namely illumination changes, clutter, occlusions or out-field of sight. Figure 1 shows snapshots of a typical run for IIDMOT involving sporadic disappearances of some body parts. For each frame, the template depicts the projection of the MMSE estimate for each ellipsoid. The IIDMOT strategy, by drawing some particles according to the detector output, permits automatic re-initialization and aids recovery after loss of observability.

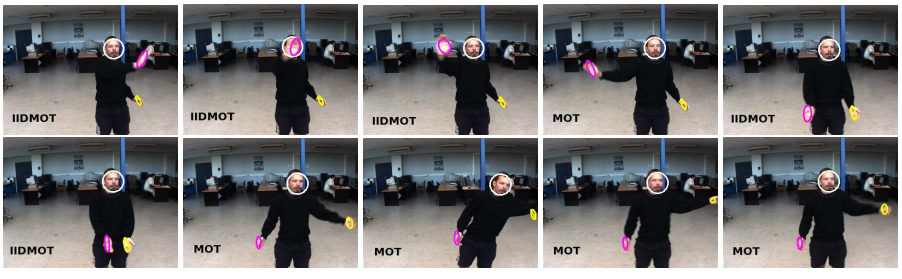

Fig. 1. Tracking scenario involving occlusion and out-field of sight with IIDMOT

Quantitative performance evaluation have 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. No tracker associated with one of the target 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 performance using multiple

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

independent particle filters (MIPF) [4], conventional IDMOT [12] strategy, and our IIDMOT strategy with data fusion.

Our IIDMOT strategy is shown to outperform the conventional approaches for a slight additional time consumption. The MIPF strategy suffers especially from "labeling" problem due to lacking modeling of interaction between trackers while the IDMOT strategy doesn't recover the target after transient loss. These results have been obtained for the "optimal" tracker parameter values listed in Table 3.

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 (3)	0.4
K	number of iterations in PF algorithm	4
d_{TH}	Euclidian distance between particles in PF algorithm	0.5
-	image resolution	256×192
-	colorspace for skin-color segmentation	CIE Lab
N_p	number of points along the ellipsoid contours	20
σ_s	standard in likelihood (5)	36
ρ	penalty in equation (5)	0.12
(σ_1, β_1)	coeff. in the repulsion "weight" (1)	$(0.12, 1.33)$
$(\sigma_{21}, \sigma_{22}, \beta_2)$	coeff. in the inertia "weight" (2)	$(1.57, 0.2, 2.0)$
Λ	standard deviation in random walk models	$\begin{pmatrix} 0.07 & 0.07 & 0.07 \\ 0.03 & 0.03 & 0.03 \\ 0.17 & 0.17 & 0.17 \end{pmatrix}$

3 Multimodal System Setup Embedded on the Robot Companion

This section gives some considerations about the integration of the above components in the architecture of our robot, depicts the execution of a target scenario in order to highlight the relevance and the complementarity of this visual component with the one dedicated to verbal and multimodal communication.

3.1 Characteristics of the Robot

Our robot is especially equipped with a 6-DOF arm, a pan-tilt stereo system on a mast, a microphone (hold by the user and cable wire connected to the robot for this first experiment), two laser scanners (figure 2-left-). From these sensors and actuators, the robot has been endowed with a set of basic functions that allows us to carry out scenarios

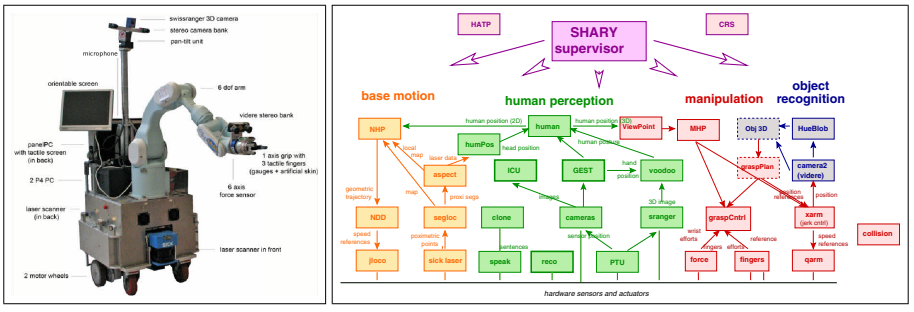

Fig. 2. The robot Jido and its layered software architecture

based on a multimodal interaction between a person and a robot, as the scenario presented in this section. Thanks to these functions, the robot is able to navigate in its environment and to recognize objects. Preliminary developments related to its vision-based functions have concerned face identification implemented through the ICU[2] module (see [11] for more details). Then, a module for gesture tracking, called GEST, has been added. It is based on the method we have described in section 2. The verbal communication mode, that deals with recognition and understanding of the user utterances, is handled by means of a dedicated module called RECO and is briefly presented below. These modules have been integrated on the robot software architecture that relies on sets of communicating modules running under the control of the platform supervisor [2].

3.2 Enabling Verbal and Multimodal Communication and Multimodal Communication

Natural communication between a person and a robot companion requires to recognize speech once uttered by the user and then to understand its meaning in relation to the current context represented by a specific task, a place, an object, an action, a set of objects or some other people involved and in some case a complementary gesture. This is the role of the RECO module integrated on the platform. Only outline and examples of results related to the type of scenarios we want to carry out are presented here.

Speech recognition: To process French utterances, we use a grammar-based speech engine, called Julian (version of the open source engine Julius developed by the Continuous Speech Recognition Consortium [1]). This engine requires essential linguistic resources : a set of acoustic models for French phonetic units (39 models, a lexicon (246 words and 428 pronunciations corresponding to phoneme sequences) drawn up from the French lexical database BDLEX [9], a set of grammars specifically designed to describe sentences related to the subtasks taken into account in our multimodal interaction scenarios : user introducing him/herself, *"Hi Jodo I'm Paul"*, giving basic movement order, *"Turn left"*, or guidance request *"Take me to the hall"*, using *"Please come here"* and other request for object exchange *"Give me this bottle"*, ... They represent 2334 different well-formed sentences, to enable communication with speech and gesture.

[2] For the acronym of "I see you".

Speech interpretation: The second part of the RECO module is dedicated to the extraction of the semantic units, directly from the recognizer output. A semantic lexicon has been designed to give the appropriate meaning of relevant words. Some are related to actions while others are related to objects or their own attributes like color or size as well as location or robot configuration parameters (speed, rotation, distance). At last, the global interpretation of the recognized utterance is transformed into a command. To be considered as valid and sent to the robot supervisor in order to be executed this command must be compatible with one of our 31 interpretation models. From the lexicon available at present, 328 interpretations can be possibly generated.

First results on the platform: Without any adaptation step, 77.6% of the 250 utterances processed have been transformed in the right command. Though, recognition and interpretation must be improved, the robot is now endowed with some abilities to interpret the user's verbal message given to him.

Fig. 3. From top-left to bottom-right : GEST (or ICU) module -left-, virtual 3D scene (yellow cubes represent hands) -middle-, current H/R situation -right-

Fusion with gesture: Deictic can be interpreted by our RECO module. If the object or location designation is precise enough (*"Put the bottle on the table"*) the right parameters are extracted from the sentence (what ? object = bottle ; where ? location = on table) and the underlying command can be generated (put(object(bottle), location(on table))). But in deictic case (*"Put the bottle there"*), the sentence analysis will mark the interpretation as "must be completed by the gesture result" and a late and hierarchical fusion strategy will be applied (put(object(bottle), location(Gesture_result)). In the same manner, for other human-dependent commands such as (*"come on my left-side"*) the same kind of strategie will be applied.

3.3 Target Robot Scenario and Preliminary Experiments

The target scenario focuses on the understanding of natural human peer-to-peer multimodal interaction in a household situation. It depends on the beforehand identification of the robot interlocutor before this one be granted permission to interact with it.

Given both verbal and gesture commands, the identified interlocutor is allowed to make the robot change its position in the environment and/or simply marks some objects the robot must catch and carry,... Figure 3 illustrates a typical run of this scenario where the robot user, after introducing himself or herself, sequences the following commands: *"go to the table", "take this bottle", "bring it to me", "go over there"*. For each step, the left subfigure shows the tracking results while the right one depicts the current H/R situation and the middle one represents the virtual H/R configuration in space (thanks to the outcome of the GEST module). The entire video and more illustrations are available at the URL `www.laas.fr/~bburger/`.

4 Conclusion

This paper presents a fully automatic distributed approach for tracking two-handed gestures and head tracking in 3D. Two lines of investigations have been pursued. First, the conventional IDMOT strategy, extended to the 3D tracking of two-handed gestures, is endowed with the nice properties of ICONDENSATION and data fusion. The amended particle filtering strategy allows to recover automatically from transient target loss while data fusion principle is shown to improve the tracker versatility and robustness to clutter. The second contribution concerns the merge of the tracker with a continuous speech interpretation process in order to specify parameters of location references and object/person IDs in verbal statements. All the components have been integrated on a mobile platform while a target robot scenario highlights the relevance and the complementarity of verbal and non verbal communication for the detection and interpretation of deictic actions during a natural peer-to-peer H/R interaction.

These preliminary robotic experiments are promising even if quantitative performance evaluations still needs to be carried out. These evaluations are expected to highlight the robot capacity to succeed in performing multimodal interaction. Further investigations will be also to estimate the head orientation as additional features in the gesture characterization. Our robotic experiments report strongly evidence that person tend to

look at pointing targets when performing such gestures. Finally, dedicated HMM-based classifiers will be developed to filter more efficiently pointing gestures.

Acknowledgements. The work described in this paper was partially conducted within the EU Projects COGNIRON ("The Cognitive Robot Companion" - www.cogniron.org) and CommRob ("Advanced Robot behaviour and high-level multimodal communication" - www.commrob.eu) under contracts FP6-IST-002020 and FP6-IST-045441.

References

1. Kawahara, T., Lee, A., Shikano, K.: Julius — an open source real-time large vocabulary recognition engine. In: European Conference on Speech Communication and Technology (EUROSPEECH), pp. 1691–1694 (2001)
2. Clodic, A., Montreuil, V., Alami, R., Chatila, R.: A decisional framework for autonomous robots interacting with humans. In: IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN) (2005)
3. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. Robotics and Autonomous Systems 42, 143–166 (2003)
4. Isard, M., Blake, A.: CONDENSATION – conditional density propagation for visual tracking. Int. Journal on Computer Vision 29(1), 5–28 (1998)
5. Isard, M., Blake, A.: I-CONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. In: European Conf. on Computer Vision, 1998, pp. 893–908 (1998)
6. Isard, M., Blake, A.: BraMBLe: a bayesian multiple blob tracker. In: Int. Conf. on Computer Vision, Vancouver, pp. 34–41 (2001)
7. Maas, J., Spexard, T., Fritsch, J., Wrede, B., Sagerer, G.: A multi-modal topic tracker for improved human-robot interaction. In: Int. Symp. on Robot and Human Interactive Communication, Hatfield (September 2006)
8. Nickel, K., Stiefehagen, R.: Visual recognition of pointing gestures for human-robot interaction. Image and Vision Computing 3(12), 1875–1884 (2006)
9. Pérennou, G., de Calmès, M.: MHATLex: Lexical resources for modelling the french pronunciation. In: Int. Conf. on Language Resources and Evaluations, Athens, June 2000, pp. 257–264 (2000)
10. Rogalla, O., Ehrenmann, M., Zollner, R., Becher, R., Dillman, R.: Advanced in human-robot interaction. In: Using gesture and speech control for commanding a robot., vol. 14, Springer, Heidelberg (2004)
11. Lerasle, F., Germa, T., Brèthes, L., Simon, T.: Data fusion and eigenface based tracking dedicated to a tour-guide robot. In: Int. Conf. on Computer Vision Systems (2007)
12. Wei, Q., Schonfeld, D., Mohamed, M.: Real-time interactively distributed multi-object tracking using a magnetic-inertia potential model. In: Int. Conf. on Computer Vision, Beijing, October 2005, pp. 535–540 (2005)