2008 IEEE/RSJ International Conference on Intelligent Robots and Systems
Acropolis Convention Center
Nice, France, Sept, 22-26, 2008

Mutual assistance between speech and vision for human-robot interaction

Brice Burger[†‡¶], Frédéric Lerasle[†¶], Isabelle Ferrané[‡¶], Aurélie Clodic[†]
[†] CNRS; LAAS; 7 avenue du Colonel Roche, 31077 Toulouse Cedex, France
[‡] IRIT; 118 route de Narbonne, 31062 Toulouse Cedex, France
[¶] Université de Toulouse; UPS; LAAS-CNRS : F-31077 Toulouse, France
E-mail: {bburger, lerasle, aclodic}@laas.fr, ferrane@irit.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. First, we propose a multiple object visual tracker which is interactively distributed and dedicated to two-handed gestures and head location in 3D. An on-board speech understanding system is also developed in order to process deictic and anaphoric utterances. Characteristics and performances for each of the two components are presented. Finally, integration and experiments on a robot companion highlight the relevance and complementarity of our multimodal interface. Outlook to future work is finally discussed.**

keywords: multiple object tracking, speech understanding, multimodal interaction, assistance robotic.

I. Introduction and framework

As the number of senior citizens increases, more research efforts have been made to develop socially interactive household robots. This field of robotics is a deep challenge because robots moving out of laboratories have to gain more social skills to improve natural peer-to-peer interaction with a novice user in his/her daily life. As speech is the most prominent communication channel for humans, a considerable number of robot assistants embed advanced speech recognition system [4]. This is not enough to realize a user-friendly interface as we, humans, omit, abbreviate and underspecify things in our utterances, that are supposed to be obtained by vision. Only few research work addresses the development of such appropriate multimodal interfaces [10]. On one hand, the mutual assistance between the speech and vision capabilities of the robot, permits to specify parameters related to person/object IDs or location references in verbal statements. On the other hand, fusing auditive and visual features are supposed to be more robust to noisy/cluttered environments than using one single feature.

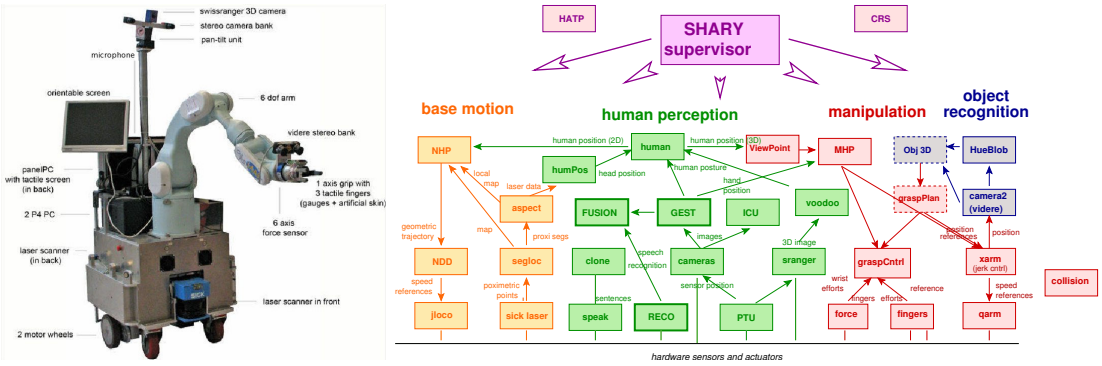
To complete/verify the message conveyed by the verbal communication channel, these interfaces consider vision techniques in order to: (i) characterize the robot surroundings *i.e.* places [6] or objects [13], (ii) perceive the human user [6], [9], [11], [13] *i.e.* his/her gestures and nonverbal reactive body motions. Besides image-based approaches [6], [9], [13], 3D positions of the user's head and hands are particularly useful, in combination with speech recognition, to specify parameters of location in verbal statements *e.g. "look here"* or *"give this object to me"*. Following [7], a first issue concerns the design of body and gesture tracker suited for gesture interpretation. This tracker has been extended in three ways. First, we propose an interactively distributed multiple object visual tracker dedicated to two-handed gestures and head location in 3D. Secondly, the tracker has been endowed with visual data fusion and automatic re-initialization. All this makes our tracker work under a wide range of viewing conditions and aid recovery from transient tracking failures due to the robot's motion or temporarily loss of observability when performing gestures. Finally, their combination with deictic and anaphoric utterances have been tested in household robotics operation with promising results. Here, gesture is used as an essential complementary information. Gesture detection could also help to reinforce communication in case of speech recognition errors.

The paper is organized as follows. Section II describes our robot companion Jido, outlines its embedded multimodal interfaces, and the target scenario we address. Section III presents the binocular tracking of the user's head and two-handed gestures in order to interpret symbolic and deictic gestures thanks to Hidden Markov Models. Section IV depicts the system dedicated to verbal communication between our robot companion and humans. Section V presents robotic experiments involving these two components. Last, section VI summarizes our contributions and discusses future extensions.

II. Jido and its target scenario

Our multimodal interface is embedded on a robot companion called Jido which is equipped with a 6-DOF arm, a pan-tilt stereo system at the top of a mast, two laser scanners (Figure 1(a)) while the human wears a wireless headset microphone. All these devices enable Jido to act as a robot assistant as it is endowed with basic functions enabling to: (i) navigate in its environment, (ii) recognize and grasp objects, (iii) detect, localize humans in its vicinity, (iv) interpret speech utterances and some gestures. All the embedded functions are managed thanks to the "LAAS" layered software architecture (Figure 1(b)) and detailed in [2].

Besides environment perception abilities, the multimodal interface has been undertaken within the demonstration scenario. This is a household situation in which Jido executes human-friendly collaborative tasks (coordinated displacements and object exchange) ordered by its disabled user. Given both verbal and gesture commands, this person



(a) (b)

Fig. 1. Jido and its software architecture.

is allowed to make the robot change its position in the environment, marks some objects the robot must catch and carry, etc. A typical set of commands is for instance: *"come on", "take this bottle on the table", "bring it to me", "go over there"*. Given this scenario, the paper focuses on the multimodal components (Figure 1(b)), namely GEST, RECO and FUSION respectively for the visual perception of the user, the speech interpretation and the fusion from user and object perception with speech interpretation. The functionalities encapsulated in these modules are presented in the following sections.

III. Visual Perception of the Robot User

A. 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 robotic 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, 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 "labelling" problems when targets undergo partial or complete occlusion. In the vein of [12], we develop an interactively distributed MOT (IDMOT) framework which is depicted in Table I. The aim is to approximate the probability density function $p(\mathbf{x}_t^i|z_{1:t})$ of the state vector $\mathbf{x}_t^i$ for body part i at time t given the set of measurements $z_{1:t}$ and the cloud of "particles" indexed by n with likelihood -or "weight"- $\omega_t^{i,n}$. When targets do not interact on each other, the approach performs like multiple independent trackers. When they are in close proximity, magnetic and inertia likelihoods (annotated $\varphi_1(.)$ and $\varphi_2(.)$) are added in each filter to handle the aforementioned problems (see [12] for more details). Our IDMOT particle filter is improved and extended in three ways. First, the conventional CONDENSATION [5] strategy is replaced by the genuine ICONDENSATION one whose importance function $q(.)$ in step 3 permits automatic (re)-initialization when the targeted human body parts appear or re-appear 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 $\alpha, \beta \in [0;1]$

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

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 $p(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i)$, 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 some consideration about the measurement function $p(z_t^i|\mathbf{x}_t^i)$ are given here below. Recall that α percent of the particles are sampled from detector $\pi(.)$ (equation (1)). These are also drawn from Gaussian distribution for head or hand configuration deduced from skin color blob segmentation in the stereo video stream.

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

TABLE I

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

Fig. 2. Tracking scenario involving occlusion and out-field of sight with our IIDMOT filter.

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. The measurement function fuses skin color information but also motion and shape cues (see [3] for more details).

Prior to their integration on Jido, 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 2 shows snapshots of a typical run involving sporadic disappearances of the hands. For each frame, the template depicts the projection of the MMSE estimate for each ellipsoid. Our strategy, by drawing some particles according to the detector output, permits automatic re-initialization and aids recovery after transient loss.

TABLE II

QUANTITATIVE PERFORMANCE AND SPEED COMPARISONS.

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

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 II presents the performance using multiple independent particle filters (MIPF) [5], conventional IDMOT [12] strategy, and our IIDMOT strategy with data fusion. The experiments, performed on an on-board 3 GHz Pentium PC, consider 100 particles to track each body part. Our IIDMOT strategy is shown to outperform the conventional approaches for a slight additional time consumption. The MIPF strategy suffers especially from "labelling" problem due to lacking modeling of interaction between trackers while the IDMOT strategy doesn't recover the target after transient loss.

B. Gesture interpretation

Gesture interpretation is reported briefly as this is not our key research goal while associated evaluations are currently performed. The typical motions pattern of eight reference gestures are classically modeled by dedicated HMMs. Five gestures serve for deictic references depending on the hold hand and the coarse pointed direction. The three last ones

TABLE III

EXAMPLES OF REQUESTS INTERPRETED BY THE ROBOT.

User starting interaction by introducing himself to the robot	"Hi robot X I'm Paul"
Basic movement orders / more advanced movement including deictic	"Turn left" / "Come here"
Guidance request in the human environment	"take me to the reception"
Interaction for object exchange including anaphora	"Give me this bottle"
Agreement / disagreement / thanks	"yes" / "no" / "thank you"

correspond to symbolic gestures, namely *"stop"* and *"come on"*[2]. The two-hands features used as the observations are derived from the tracked head position while each HMM has been found to perform best with 3-state model each. Preliminary evaluations on sequence dataset issued from a commercial human motion capture highlights that our gesture recognizer scored at about 91% sensitivity and 92% selectivity. Evaluations dedicated to our IIDMOT multi-tracker output, and so on noisy observations, will follow.

IV. SPEECH INTERPRETATION

The speech understanding system must recognize continuous speech, or even spontaneous, and must handle some linguistic phenomena ordinarily used in conversational speech and multimodal communication. We present hereafter how our robot can perceive and understand the information conveyed by the spoken message, how it can infer that a gesture event is necessary to complement speech or how gesture can strengthen speech in case of recognition errors.

A. Integration of a speech recognition module on the robot platform

To fulfil the platform architecture and software requirements, we have chosen to use a grammar-based recognizer. Julian, is a version of Julius developed by the Continuous Speech Recognition Consortium [1] which is itself an open source speech recognition engine. To process French utterances, a set of acoustic models (for phonetic units), a phonetic lexicon of words and a set of language models must be provided.

B. Linguistic resources for speech recognition

The acoustic models stem from previous work on large vocabulary speech transcription. They are HMM-based (37 phoneme and one short and one long pause, each one is a 3-state model with 32 Gaussians per state) and have been trained using the HTK toolkit on about 31 hours of Broadcast News recorded on French radios. Though speech recognition in a human-robot interaction context is a different task from the initial one, these acoustic models have not been adapted yet to this new applicative context, while the lexicon and the grammars have been specifically designed for it. The lexicon with 246 words and their different pronunciations (corresponding to 428 phoneme sequences) have been drawn up from the French lexical database BDLEX [8]. This vocabulary has been selected according to different subtasks as shown in the table III[3]. In order to focus on the multimodal aspect of human-robot communication, we will take a particular interest in recognizing and interpreting deictic and anaphora.

The language models, which are implemented through different context free grammars related to the above subtasks, describe an overall set of 2334 well-formed sentences.

C. Speech interpretation

This part of the RECO module processes speech recognition outputs in order first to extract the semantic units that are relevant in the user utterance and then to build the appropriate interpretation. It is based on a semantic lexicon specifically designed which associates relevant words with their interpretation in the context of the aforementioned subtasks. Some words are related to actions while others are related to objects, object attributes like color or size as well as location and robot configuration parameters (speed, rotation, distance). A semantic analysis step combines word semantic information and builds a global interpretation which is compared with available interpretation models. If one of them is compatible with the utterance interpretation, we consider that a valid command can be generated and sent to the robot supervisor in order to be executed.

D. Interpreting deictic and anaphora

Deictic words (here, there, ..) are defined in our semantic lexicon as related to a location which will be given by means of a gesture. This is specified by a semantic feature (location = Gesture_location?) . For example, if the verbal designation of an object or a location is precise enough (*"Put the bottle on the table"*) the parameters are directly extracted from the sentence according to the relevant words and their semantic information. In our semantic lexicon, the word "put" is associated with the meaning "put something somewhere" which is represented by the set of semantic features (action = put ; object = What? ; location = Where?). The sentence analysis instantiate the missing parameters (What? and Where?) and the underlying command can be generated (put(object=bottle, location = on_table)). But in deictic case (*"Put the bottle there"*), the semantic analysis will mark the interpretation as "must be completed by the gesture result" and a late and hierarchical fusion strategy will be applied to complete the command that has been generated (put(object=bottle, location=Gesture_location?)) (see section V). In the case of an anaphoric sentence (*"Take this glass"* (action = take; ref_object = (object = glass ; ref_location = Gesture_location?))) and other human-dependent commands such as (*"Come on my left-side"* (action = go ; relative_location = (ref_location = User_position? ; side = left))) the same kind of strategy will be applied. For the moment, only location reference are taken into account. In a human-robot dialog prospect anaphora could also be

[2]From single or two-handed gestures.

[3]Examples are given in English for an illustration purpose.

TABLE IV

EXPERIMENTAL RESULTS OF THE SPEECH RECOGNITION COMPONENT ALONE AND OF THE GLOBAL SPEECH INTERPRETATION SYSTEM.

subtask	COR_W	ACC	COR_S	COR_COM
starting/closing interaction	88.34%	81.97%	67.19%	71.88%
basic movement orders	89.63%	81.72%	65.10%	70.05%
basic object manipulation orders	86.41%	80.62%	62.50%	66.25%
deictic	94.79%	90.77%	82.81%	83.33%
guidance request	83.30%	78.66%	48.75%	71.25%
complete order for object exchange	86.41%	78.80%	61.25%	66.88%
anaphoric order for object exchange	85.62%	69.38%	47.92%	48.96%
agreement/disagreement	94.12%	89.34%	79.38%	83.75%
robot status	81.44%	77.34%	75.00%	75.00%
overall results	84.15%	75.84%	66.19%	71.69%

solved by means of an history, which is not taken into account yet.

E. First evaluation of speech recognition and interpretation

In order to evaluate the RECO module, a list of 50 well-formed sentences related to the different tasks described above has been drawn up. Each one has been uttered 32 times so our first evaluation corpus counts 1600 utterances. Fourteen different speakers were involved in these experiments. These first results are given in the table IV : percentages of correct words (COR_W), accuracy at word level (ACC), correct sentences (COR_S) and correct commands (COR_COM). A command has been generated from each valid interpretation of a speech recognition result and then compared with the corresponding reference command.

General comments can be made about these results. For each subtask, COR_COM is greater than COR_S (or equal in the last case). The speech recognition errors, at the word level, have less impact on the command than on the sentence. If a word omitted, inserted or substituted by another one, is not semantically relevant, this will not have a real impact on the command generation, but the sentence will be considered as completely wrong. This explains the COR_COM higher rates. The results for deictic orders are correct unlike the anaphoric ones, especially for the sentence (*"Take this"*). Only the best recognition output is taken into account at the moment. At mid-term, the N-best results will be considered at the fusion level. If a gesture has been interpreted, and if the recognized sentence does not need a complementary gesture, we can detect an incoherence and we could propose another interpretation. Further developments will consider such a multiple hypothesis strategy while the acoustic models will be adapted to the robotic context.

V. THE MULTIMODAL INTERFACE AND LIVE EXPERIMENTS

A. Vision and audio fusion

Vision and audio data are merged using a rule based approach. The speech is used as the main channel : the RECO module, thanks to its semantic knowledge, identifies actions needing a gesture disambiguation. Vision is used in a late and hierarchical fusion strategy to complete this input information.

For deictic commands, like *"put the bottle there"*, and its interpretation (put(object=bottle, location=Gesture_location?) the non instantiated parameters (here, the bottle position) are specified by the FUSION module via the line of sight between head and the hold hand extracted by the GEST module. In these cases, we assume that we can use head and hands 3D positions at the end of the speech utterance to extract the pointed direction, knowing that speech and gestures are strongly correlated in time. For human-dependent commands such as *"come on my left-side"* and its interpretation (action = go ; relative_location = (ref_location = User_position? ; side = left))) , the same kind of strategy is applied, extracting the human position from the head location.

B. Live experiments

The integration of the multimodal interface on Jido enables us to perform online experiments in our lab. Figure 3 illustrates a typical run of the scenario. For each step, the main picture depicts the current H/R situation, while the subfigure shows the tracking results of the GEST module.

The robot succeeds to interpret a sequence of commands by melting multimodal features in the FUSION module. The entire video and more illustrations are available at the URL `www.laas.fr/~bburger`.

More globally, the robot succeeds to execute the scenario in the majority of runs with Jido successfully bringing the bottle to its human user. The principal failures are attributable to the precision of pointing gesture which decreases with the angle between the head-hand line and the table. The multimodal interface is shown to be robust enough to allow continuous operation for the long-term experimentations that are intented to be performed.

VI. CONCLUSION

In this paper, we propose a scenario for Human-Robot interaction based on mutual assistance between speech and vision which rely on three modules integrated on a robotic platform. Before integration on the platform, each module and the underlying methods implemented are described, followed by some results provided by a step of quantitative evaluation of the module performances. The first contribution describes a fully automatic distributed approach for tracking two-handed gestures and head tracking in 3D. 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. Speech recognition and interpretation constitutes the

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

second contribution, focusing on the interpretation of utterances related to predefined subtasks and more particularly on deictic and anaphoric commands requiring fusion with gesture events. Then, in order to specify parameters for location references and object/person IDs and complement verbal statements, we present the outlines of the late fusion performed from both speech and gesture analysis. As shown by the scenario execution, these preliminary robotic experiments are promising even if speech recognition performances must be improved and quantitative performance evaluations still need to be carried out. These evaluations are expected to highlight the robot capacity to succeed in performing multimodal interaction. Further investigations will also be to : (i) process more natural and flexible utterances about object manipulation tasks, (ii) 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. Dedicated HMM-based classifiers will be developed to filter more efficiently pointing gestures. Another investigation line will be to study other fusion methods based on the conjoint modelling of speech and gesture.

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 Future and FP6-IST-045441.

REFERENCES

[1] T. Kawahara A. Lee and K. Shikano. Julius — an open source real-time large vocabulary recognition engine. In *European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1691–1694, 2001.

[2] R. Alami, R. Chatila, S. Fleury, and F. Ingrand. An architecture for autonomy. *International Journal of Robotic Research (IJRR'98)*, 17(4):315–337, 1998.

[3] Brice Burger, Isabelle Ferrané, and Frédéric Lerasle. Multimodal interaction abilities for a robot companion. In *Int. Conf. on Computer Vision Systems (ICVS'08)*, pages 549–558, Santorini, Greece, 2008.

[4] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42:143–166, 2003.

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

[6] J. Maas, T. Spexard, J. Fritsch, B. Wrede, and G. Sagerer. A multi-modal topic tracker for improved human-robot interaction. In *Int. Symp. on Robot and Human Interactive Communication*, Hatfield, September 2006.

[7] K. Nickel and R. Stiefehagen. Visual recognition of pointing gestures for human-robot interaction. *Image and Vision Computing (IVC'06)*, 3(12):1875–1884, 2006.

[8] G. Pérennou and M. de Calmès. MHATLex: Lexical resources for modelling the french pronunciation. In *Int. Conf. on Language Resources and Evaluations*, pages 257–264, Athens, June 2000.

[9] O. Rogalla, M. Ehrenmann, R. Zollner, R. Becher, and R. Dillman. *Advanced in human-robot interaction*, volume 14, chapter Using gesture and speech control for commanding a robot. Springer-Verlag, 2004.

[10] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, and W. Adams. Spatial language for human-robot dialogs. *Journal of Systems, Man, and Cybernetics*, 2(34):154–167, 2004.

[11] R. Stiefelhagen, C. Fügen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel. Natural human-robot interaction using speech head pose and gestures. In *Int. Conf. on Intelligent Robots and Systems (IROS'04)*, Sendal, October 2004.

[12] Q. Wei, D. Schonfeld, and M. Mohamed. Real-time interactively distributed multi-object tracking using a magnetic-inertia potential model. In *Int. Conf. on Computer Vision (ICCV'05)*, pages 535–540, Beijing, October 2005.

[13] M. Yoshizaki, Y. Kuno, and A. Nakamura. Mutual assistance between speech and vision for human-robot interface. In *Int. Conf. on Intelligent Robots and Systems (IROS'02)*, pages 1308–1313, 2002.