Towards Shared Attention through Geometric Reasoning for Human Robot Interaction

Luis F. Marin-Urias, Emrah Akin Sisbot, Amit Kumar Pandey, Riichiro Tadakuma and Rachid Alami

Abstract— **Human Robot Interaction brings new challenges to the geometric reasoning and space sharing. The robot should not only reason on its own capacities but also consider the actual situation by looking from human's eyes, thus "putting itself to human's perspective".**

In humans, the "visual perspective taking" ability begins to appear by 24 months of age and is used to determine if another person can see an object or not. In this paper, we present a geometric reasoning mechanism that employs psychological concepts of "perspective taking" and "mental rotation" in order to reason what the human sees, what the robot sees and where the robot should focus to share human's attention. This geometric reasoning mechanism is demonstrated with HRP-2 humanoid robot in a human-robot face-to-face interaction context.

I. INTRODUCTION

As the research on human-robot interaction approaches more and more towards robot interacting closely with people having knowledge, beliefs, attention, perspectives and capacities, the necessity of taking these notions into account in robot's design appeared. In this paper, we are focusing on geometric tools to model the attention of the human towards the goal of enabling the robot to share it.

In [1], the notion of "joint attention" is defined as a bilateral process between at least two agents (human and/or robot), where both are aware about the intentions of each other. This process has to contain at least four prerequisites:

- Attention detection. To follow other agent's attentional behavior (i.e. Gaze following)
- Attention manipulation. To influence the attentional behavior of the other agents.
- Social coordination. Joint coordinated actions (turn-taking, role-switching, etc.)
- Intentional understanding. Agents must notice if they share the same intention to achieve the same goal.

Attention sharing requires, among others, the notions of "perspective taking" and "mental rotation" taken into account robots reasoning mechanism. "Mental rotation" is the ability to acquire the representation of the environment from another point of view. On the other hand, "perspective taking" is the general notion of reasoning from another person's point of view to obtain a representation of that person's knowledge. In the context of this paper, we are interested in mental rotation and visual perspective taking where the robot should place itself to human's place to determine what he is actually seeing. Visual perspective taking is one of the elements that contribute to the general knowledge and actual attention of a person because of the fact that a thing that is seen becomes automatically known.

The authors are with CNRS; LAAS, 7 Avenue de Colonel Roche, 31077, Toulouse France and Université de Toulouse: UPS, INSA, INP, INSAE, LAAS, F-31077, Toulouse, France.

In literature, the research towards joint attention is mainly based on detecting human gaze and following it in an image to find out what the human is looking and focusing the robot towards that objective [3], [4]. Visual occlusions are not taken into account and the reasoning done is generally on the 2D image of the robot.

[5] describes that the imitation of some other "human social cues" has to be taken into account (added to tracking gaze system). It is mentioned that the recognition or execution of a gesture that can manipulate the attention (i.e. declarative and imperative pointing and "eye contact") of an attentional partner, helps to the development of better social behaviors. The implementation of what is called "mutual gaze" has shown an efficient task-based decomposition to achieve Joint attention.

Imai et al. presented a robot platform [2] that performs vocal utterances added to "eye contact" with predefined pointing gestures that carries the attention of the human to an referenced object (performing joint attention). A basic 3D reasoning is employed to infer the position of the pointed object.

In [6], Brooks et al. presented Leonardo showing also some characteristics of shared attention on the robot gestures to communicate with human. The robot looking at the same button, that human is making reference, included reasoning on human's pointing motion.

An important geometric tool for attention sharing is the mental rotation. This notion is often used in computer graphics and simulation where the screen allows the watcher to see a scene from different angles (e.g. to simulate human-like view in order to obtain fast object drawing [10], [11], or to increase the level reality of environment, home or car design [12]).

Mental rotation is also used in robotics to simulate robot sensors [14] and also in the interface design of games and human-computer interaction scenarios [15]. The most common use of mental rotation in HRI is the research on teleoperation where the operator needs to see the scene that the robot actually sees [13].

Perspective taking is also employed in HRI where the robot needs to put itself on human's place and share his state of mind. In [8], Breazeal et al. presented a learning algorithm that takes into account information about teacher's visual perspective with a predefined perspective information. Berlin et al. used a 3D simulated dynamic environment for

the same learning scheme where perspective taking entered to modify the belief system [9].

Trafton et al. presented a robot [7] that uses geometrical reasoning with perspective taking to take decisions about human reasoning. They proposed a polyscheme propositional system to take decisions in ambiguous scenarios. In a similar way, Johnson et al. [16] used the perspective taking to make a more accurate action recognition in the interaction between two robots.

In this paper, we present a geometric reasoning system involving mental rotation and perspective taking capabilities that can be used for human-robot attention sharing. This system reasons entirely in 3D, modeling different visual capabilities for each agent and determining objects from different angles with visual obstruction taken into account. This will enable the robot capability to plan motions in order to attempt sensor-based configurations, where the robot can share or influence the human attention, as shown in previous work in [20], [19].

II. GEOMETRIC TOOLS

Attention sharing requires psychological notions of perspective taking and mental rotation taken into account in robots reasoning. As mentioned in previous section, perspective taking is the general notion of taking another person's point of view to acquire an accurate representation of that person's knowledge. In the context of this paper, we are interested in visual perspective taking where the robot should place itself to human's place to determine what he is actually seeing.

A. Mental Rotation

In order to find out what human is seeing, we attach a virtual camera into "human's eyes" in its model within Move3D [17] simulation and planning environment. The attached camera will move as the configuration of the human model changes. To determine what is perceived by this camera, we use 2D perspective projection of the 3D environment. This projection is obtained from an image taken from the human's eyes point of view.

The obtained result is the matrix $MatP$ where the value of the position (x, y) represents one point in the projection image in human's field of view. A 2D projection of the scenario shown on figure 1 is illustrated in the figure 2.

The 2D projection image, which is the result of this mental rotation process, represents the points of the environment that the human is actually seeing.

Even though the information on visible and invisible points is interesting, for an HRI scenario the most important information that can be extracted from this image is, which objects, humans, obstacles or robots are actually seen by the human. This image will be used as an input of perspective taking mechanism that reasons the visibility of each body in this image.

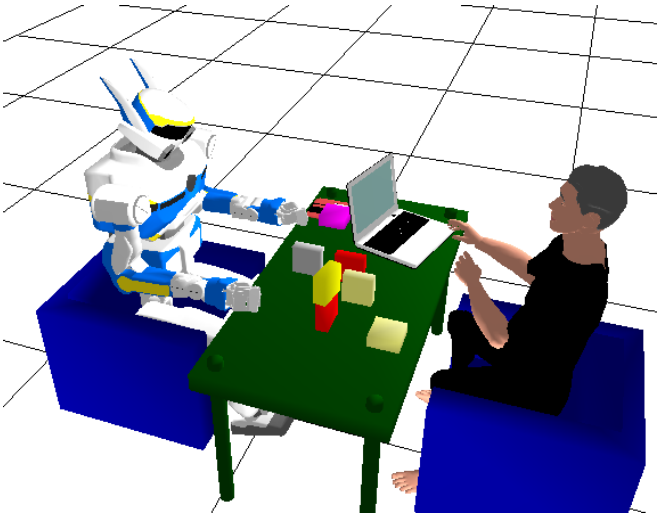

Fig. 1. A scenario where the human and the robot are sitting face to face

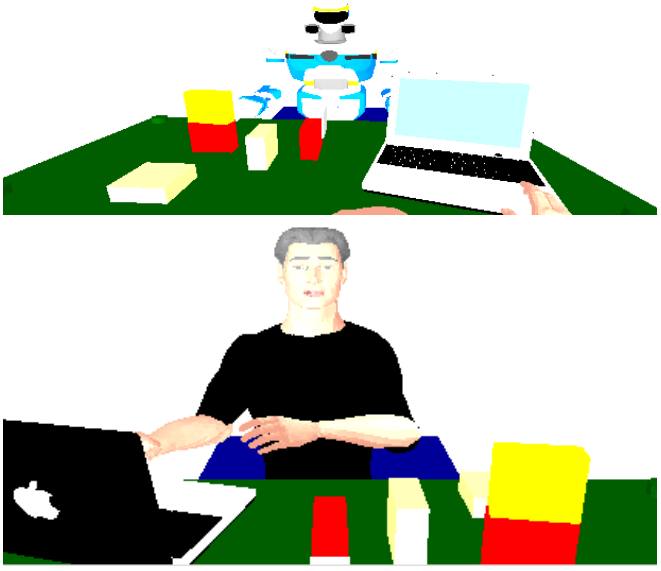

Fig. 2. Computed perception of the robot (top) and human (bottom). The perception depends on the sensor capabilities and configurations

B. Visual Perspective Taking

We define "Projection rate" Pr as the projection percentage of an element El (object, human, obstacle or robot) on the environment represented in $MatP$. Pr is obtained by:

$$Pr(El) = \Sigma MatP(x, y) \mid (x, y) \epsilon El$$

The projection rate of an element that is not projected Pr_{hidden} can be obtained with:

$$Pr_{hidden}(El) = Pr_{desired}(El) - Pr_{visible}(El)$$

where $Pr_{visible}$ is the projection rate that considers visual obstructions (only visible projection). On the other hand, $Pr_{desired}$ is the relative projection obtained without considering objects in the environment (as it should look without

visual obstacles). Figure 3 illustrates the difference between desired and visible relative projections of the robot from human's point of view. As the perspective taking system reasons the visibility by taking into account everything in the 3D environment, including the human himself, human's hand causes a small visual occlusion on the laptop (figure 4).

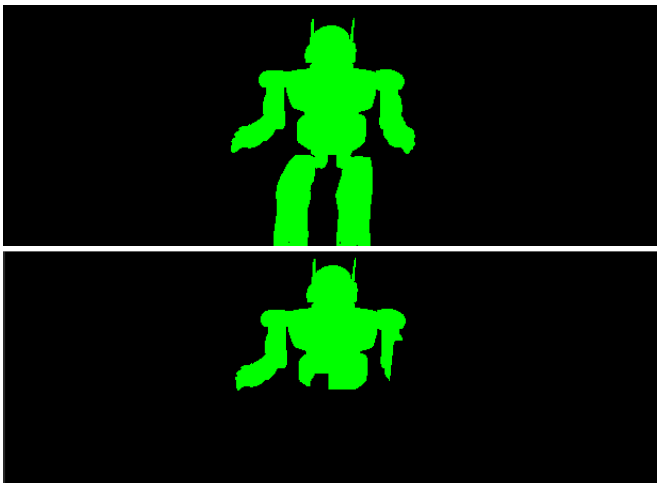

Fig. 3. Relative projections: The robot is the target and differs from other elements on the environment. a)Desired relative projection b) Visible relative projection. The table and the objects are blocking the human's view.

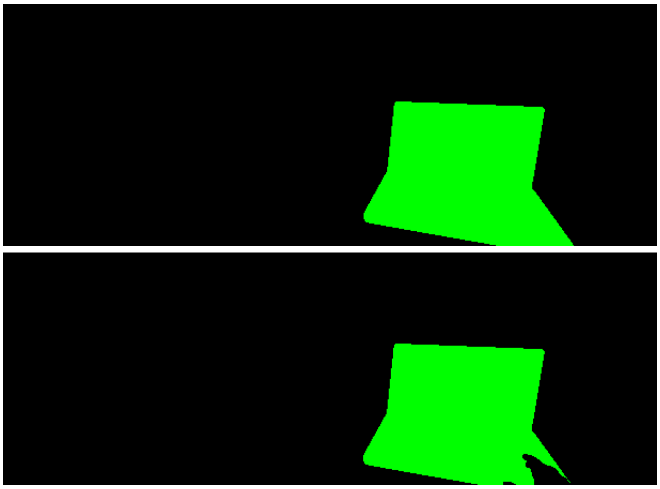

Fig. 4. Relative projections: The laptop on the table is the target and differs from other elements on the environment. As the perspective taking system reasons the visibility by taking into account everything in the 3D environment, including the human himself, human's hand causes a small visual occlusion on the laptop.

Visibility quality percentage of an element El defined by $Watch$ is determined by:

$$Watch(El) = \frac{Pr_{visible}(El)}{Pr_{desired}(El)}$$

Finally, an element is considered visible to the human by: $Watch(El) \geq \mu$ where μ is a threshold that corresponds to a desired percentage.

A snapshot of a scenario where a person is sitting on a table is illustrated in figure 5. In this example, the human is looking at the laptop. By using mental rotation and perspective taking systems, the robot determines that the object in human focus is the laptop. Although the human looks also in the direction of the bottle and the white box, these two objects are evaluated as invisible by the system because of the occlusion of the laptop.

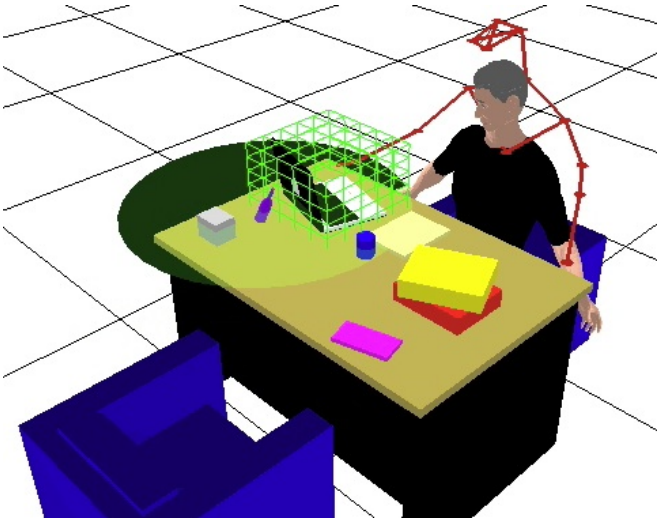

Fig. 5. An instance of the situation assessment. Object marked with a green wire box (the laptop) is evaluated as visible. The violet bottle and the white box are not visible to the human because of the occluding laptop.

C. Seen Objects

Some objects in the environment can be considered as fixed obstacles and can be excluded from the perspective taking. This functionality can allow the robot to react according to the context and human's activity. In the scenario illustrated by Figure 5, the table is considered as an obstacle and is not returned by perspective placement system. In the context of a person sitting next to a table, we can consider that the attention of the human will be mainly on the object on the table but not on the table itself.

The human is provided with a wide field of view. Nevertheless, when it is centering its attention to something the visual attention reduces its size to a particular cone form as we can see in [21]. The objects outside the attentional field of view, also can be ignored from the perspective taking process.

On the final elimination step, we pass the list of objects to the perspective process and like that we can obtain the objects perceived from the human's perspective.

Although all the process of elimination of the non attentional objects (objects that are not of interest), we still could have many objects that can enter in the attentional field of view and that are not occluded by another object. Perspective taking of the human gaze is not enough to determine an attentional object, so that we have to consider temporal constraints of attention. Human has to spend little time on an object to become an attentional object.

Finally, in the case of ambiguous attentional objects, we take the first closest object to the line in the center of the visual attention cone. In other words the closest to human's line of perception.

For acquiring more reliable data, a process of data fusion with utterances and context should be applied to the process.

III. INTEGRATION AND RESULTS

A. Scenario

The experimental environment resembles a Face to Face interaction scenario of human and our Humanoid Robot HRP2. The table serves as common work platform and the objects on the table are a toolbox, a small box and a cup. Figure 6 shows the real scenario and figure 8 shows its 3D representation in the interface of our Move3D [17] software platform, where the geometric tools are implemented.

Note that apart from using the static model of the environment, our system puts objects dynamically in this 3D model, which has been described in next section.

Fig. 6. Face to face scenario. The table is the common work platform for the interaction objects.

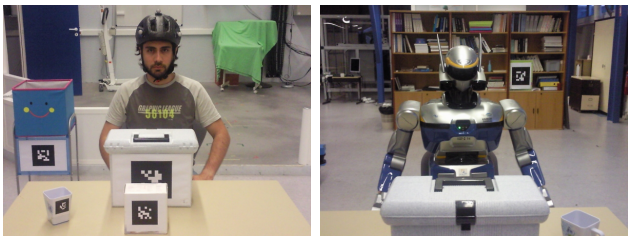

Fig. 7. The scenario from robot's and human's eyes.

B. Implementation and Results

The entire system has been carried to our HRP2 robotic platform. HRP2 is a humanoid robot developed by Kawada Industries, Inc. It has 30 degrees of freedom. In LAAS-CNRS, it has the vision system composed of four cameras on its head. The robot's height is 1570 mm and its width is 613 mm. Its mass is 58kg, including batteries.

The system architecture is consisting of various task-specific dedicated OpenGenom modules [18]. The environment Move3D is managed by one of these modules, called

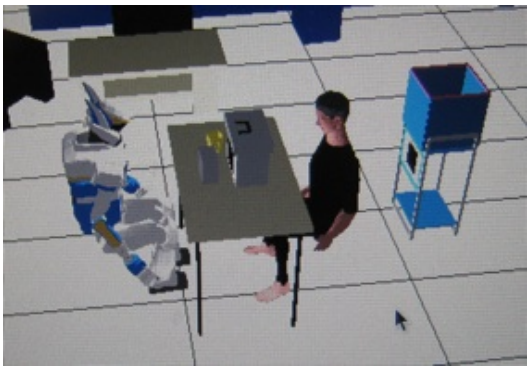

Fig. 8. The 3D representation of the same environment including the robot, the human and the objects.

GEO, in order to interface it with the other modules. A scheme of the system architecture is illustrated on the figure 9, showing all the data flow with the GEO module.

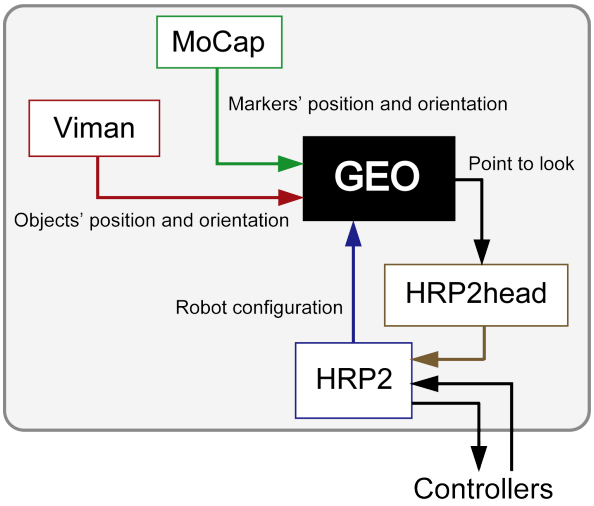

Fig. 9. The system architecture in the robot. The GEO module receives markers positions from the MoCap Client and sets the human position and head orientation. From Viman, object positions are obtained and updated. Once the reasoning about human perspective is done, robot head configuration is passed to the controller modules.

The acquisition of the dynamic changes in the environment is mainly done by two modules: the vision based module *ViMan* and head markers detection based on motion capture system. ViMan uses stereo cameras of the robot and tags on the object to identify and calculate its position in the 3D space. The GEO module continuously obtains this 3D positions and orientations, and then updates the environment placing models of the objects on the table (or elsewhere) dynamically, at the moment of its detection.

As a temporal platform for obtaining a precise motion of the human head as also the gaze orientation, a Motion Capture system was installed in the experimental environment. It consists of 10 cameras at different positions which covers a volume in the environment.

The person whom the robot is intended to interact, is equipped with an special cap consisting of a set of markers.

Since the motion capture system can only localize and track human head position and orientation, the gaze direction of the human is simplified to his head direction. With a more precise "gaze detection", the overall system can be easily adapted to human eye motions.

The server of the motion capture system broadcasts the position of the markers, a dedicated client running communicates to the server and updates the position of the markers. Our GEO module acquires data from this client, interprets markers position and geometrically calculates the orientation of the human head in real time.

Once the perspective taking process is done and the attentional object has been defined, the robot turns its head to the center of gravity on this object.

Figure 10 shows the simplest face to face scenario. It illustrates the results on an image sequence from different videos, here the robot turns the head each time it detects that the human is changing of attentional object. The attentional object can be verified because it is on the center of the image of the robot's camera.

Figure 11 illustrates the results when two objects are in the same field of view but one of them is occluding the other. The attentional object is the one that the human can see, and not the hidden one [1].

IV. CONCLUSION AND FUTURE WORK

In this work we have presented a first step of the development of a set of useful geometric tools that helps to the development of a shared attention ability in human robot interaction.

We also have shown the importance of the implementation of algorithms based on perspective taking and mental rotation concepts to obtain visual attentional objects.

Furthermore, we have not only developed the geometric tools on a 3D simulation environments but also we have shown its implementation on a humanoid robotic platform, and obtaining promising results. Nevertheless, we are still working on the evolution and improvement of the system.

For this paper, we have activated two joints on the neck of the robot. This means that, at this step of the integration of the geometric tools, it only work on a static behavior in a face to face interaction, moving only the robot's head to look at the same object that the human is looking at.

The system is intended to perform more complex shared attention and interaction actions. In the very near future, this system will be able to plan sensor based configurations and motions using its arms, hand, and waist joints, based on Inverse kinematics and collision detection. All this, integrating it with other systems of human aware motion planning that also use perspective taking and mental rotation concepts [20].

In order to evaluate the overall system we also plan to conduct user studies with naive users using the simulation environment as well as the robot.

Also, we are currently working on process of influence of the human visual attention, reasoning human perspective and tracking the human gaze on objects that the robot is currently manipulating. All this to perform human understandable actions to achieve joint activities of interaction.

[1]The videos of the mentioned results are shown in http://www.laas.fr/∼lfmarin/Videos.html

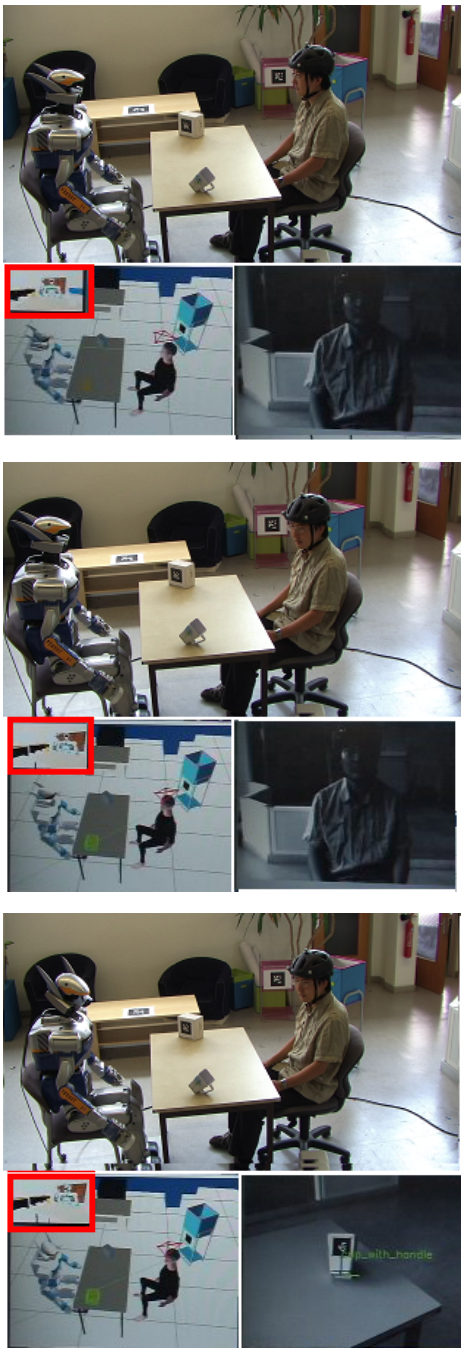

Fig. 10. Scenario 1: the robot looks to the object that the human is looking. In the images: The scenario (up), the GEO-Move3D interface showing the process and the attentional object marked with a green grid box around the object (down-left) and robot's camera (down-right)

V. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Information and Communication Technologies Seventh Framework Programme FP7/2007-2013 under grant agreement no 215805, the

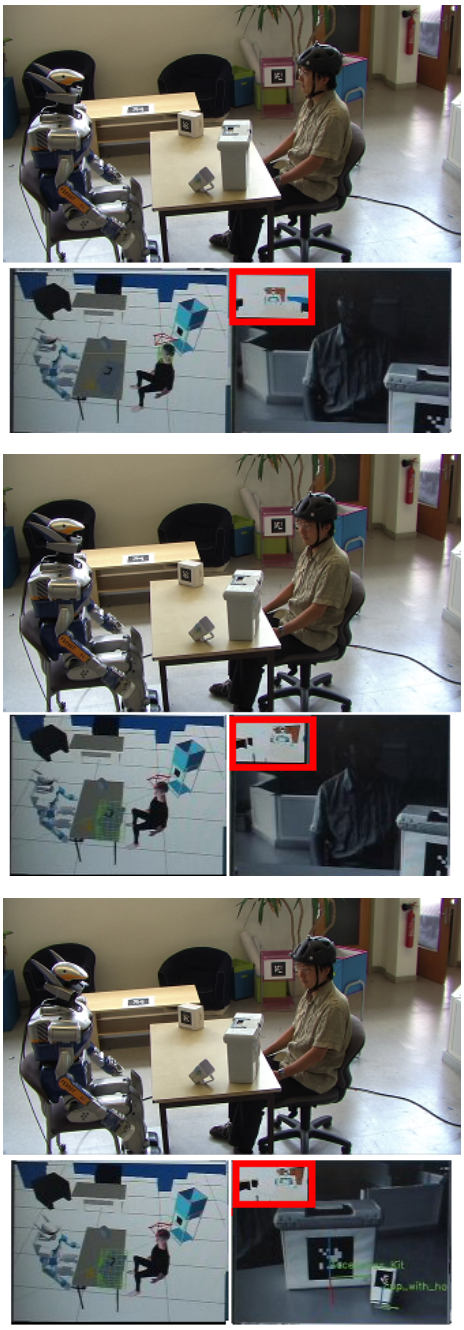

Fig. 11. Scenario 2: the robot is capable of detecting and looking at the object of attention, avoiding visual occlusions between objects. The toolbox is occluding the small cup to the human.

CHRIS project. It is also partially supported by ANR in the framework of AMORCES Project (grant ANR-07-ROBO0004).

REFERENCES

[1] Kaplan, F. and Hafner, V.V, *The challenges of joint attention* The challenges of joint attention, Interaction Studies, 7 (2): 135-169. (2006)

[2] Imai, M.; Ono, T.; Ishiguro, H., *Physical relation and expression: joint attention for human-robot interaction*, Industrial Electronics, IEEE Transactions on Volume 50, Issue 4, Aug. 2003 Page(s):636 - 643

[3] Han-Pang Huang; Jia-Hong Chen; Hung-Jing Jian, *Development of the joint attention with a new face tracking method for multiple people* , Advanced robotics and Its Social Impacts, 2008. ARSO 2008. IEEE Workshop on 23-25 Aug. 2008 Page(s):1 - 6

[4] Yukie Nagai and Koh Hosoda and Minoru Asada, *How does an infant acquire the ability of joint attention?: A Constructive Approach*, Lund University Cognitive Studies pages 91-98 2003

[5] Scassellati, Brian, *Imitation and Mechanisms of Joint Attention: A Developmental Structure for Building Social Skills on a Humanoid Robot*, in C. Nehaniv, ed., Computation for Metaphors, Analogy and Agents, Vol. 1562 of Springer Lecture Notes in Artificial Intelligence, Springer-Verlag, 1999.

[6] Brooks, J. Gray, G. Hoffman, A. Lockerd, H. Lee and C. Breazeal. *Robot's Play: Interactive Games with Sociable Machines*, ACM Computers in Entertainment, 2(3), 1-18, 2004

[7] Trafton, J. Gregory, Cassimatis, Nicholas L., Bugajska, Magdalena D., Brock, Derek P., Mintz, Farilee, and Schultz, Alan C., *Enabling Effective Human-robot Interaction Using Perspective-taking in Robots* IEEE Transactions on Systems, Man, and Cybernetics, Part A, 460-470. 2005

[8] Breazeal, Cynthia, Berlin, Matt, Brooks, Andrew, Gray, Jesse, and Thomaz, Andrea L., *Using Perspective Taking to Learn from Ambiguous Demonstrations*, Robotics and Autonomous Systems, 385-393, 2006.

[9] Berlin, Matt, Gray, Jesse, Thomaz, Andrea L., and Breazeal, Cynthia. *Perspective Taking: An Organizing Principle for Learning in Human-Robot Interaction*, International Conf. on Artificial Intelligence, AAAI. Boston, Mt., 2006.

[10] Hansong Zhang and Dinesh Manocha and Tom Hudson and Kenny Hoff, *Visibility Culling Using Hierarchical Occlusion Map*, Proceedings of SIGGRAPH ,1997.

[11] Chhugani, Jatin, Purnomo, Burdirijant, Krishnan, Shanka, Cohen Jonatha, Venkatasubramanian, Johnson, David and Subodh, Kumar , *"vLOD : High-Fidelity Walkthrough of large Virtual environments"*, IEEE Transactions on Visualization and Computer Graphics Vol 11 No. 1 2005

[12] Pascal Mueller, Peter Wonka, Simon Haegler, Andreas Ulmer and Luc Van Gool, *Procedural Modeling of Buildings*, ACM Transactions on Graphics 2006, Vol. 25 pages 614-623 No.3

[13] M. Alejandra Menchaca-Brandan, Andrew M. Liu, Charles M. Oman and Alan Natapoff, *Influence of Perspective-taking and Mental Rotation Abilities iin Space Teleoperation*, HRI '07:Proceeding of the ACM/IEEE international conference on Human-robot interaction, 2007, Arlington, Virginia, USA.

[14] Josh Faust, Cheryl Simon and William D. Smart, *A Video Game-based Mobile Robot Simulation Environment* , Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2006 Beijing, China.

[15] Hsu, S. W., and Li, T. Y., *Third-Person Interactive Control of Humanoid with Real-Time Motion Planning Algorithm*. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2006.

[16] M. Johnson, and Y. Demiris *Perceptual Perspective Taking And Action Recognition*. International Journal of Advanced Robotic Systems,Vol.2 (4) pp 301-308, Dec. 2005.

[17] T. Siméon, JP. Laumond, F. Lamiraux, *Move3D: a generic platform for motion planning*, 4th International Symposium on Assembly and Task Planning, Japan, 2001.

[18] S. Fleury and M. Herrb and R. Chatila, *Genom: a tool for the specification and the implementation of operating modules in a distributed robot architecture* IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 1997, Grenoble, FR.

[19] E. Akin Sisbot and Luis F. Marin-Urias and Rachid Alami, *Spatial Reasoning for Human Robot Interaction*, Proc in (IEEE/RSJ) International Conference on Intelligent Robots and Systems, IROS07 2007, San Diego, CA, USA.

[20] Luis F. Marin-Urias and E. Akin Sisbot and Rachid Alami, *Geometric Tools for Perspective Taking for Human-Robot Interaction* Seventh Mexican International Conference on Artificial Intelligence 2008, Mexico City, Mexico.

[21] Notger G. Muller and Maas Mollenhauer and Alexander Rosler and Andreas Kleinschmidt, *The attentional field has a Mexican hat distribution* Journal on Vision Research vol. 45 no. 9 pp. 1129-1137, 2005.