Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, Technische Universität München, Munich, Germany, August 1-3, 2008

Joint-Action for Humans and Industrial Robots for Assembly Tasks

Claus Lenz, Suraj Nair, Markus Rickert, Alois Knoll
Robotics and Embedded Systems Lab, Department of Computer Science
{lenz,nair,rickert,knoll}@in.tum.de
Wolgang Rösel Machine Tools and Industrial Management, Department of Mechanical Engineering
wolgang.roesel@iwb.tum.de
Jürgen Gast, Alexander Bannat, Frank Wallhoff
Human-Machine Communication, Department of Electrical Engineering and Information Technologies
{gast,bannat,wallhoff}@tum.de
Technische Universität München, Germany

Abstract— **This paper presents a concept of a smart working environment designed to allow true joint-actions of humans and industrial robots. The proposed system perceives its environment with multiple sensor modalities and acts in it with an industrial robot manipulator to assemble capital goods together with a human worker. In combination with the reactive behavior of the robot, safe collaboration between the human and the robot is possible.Furthermore, the system anticipates human behavior, based on knowledge databases and decision processes, ensuring an effective collaboration between the human and robot. As a proof of concept, we introduce a use case where an arm is assembled and mounted on a robot's body.**

I. Introduction

The state-of-the-art in human-robot collaboration is mainly based on a master-slave level where the human worker teleoperates the robot or programs it off-line allowing only static tasks to be executed. To ensure safety, the workspaces of humans and robots are strictly separated in time or in space. For instance, in the automobile industry, human workers are completely excluded from the production lines where robots execute assembly steps. On the other hand, robots are not integrated in assembly line manufacturing along with human workers.

These approaches do not take advantage of the potential for humans and robots to work together as a team, where each member has the possibility to actively assume control and contribute towards solving a given task based on their capabilities. Such a mixed-initiative system supports a spectrum of control levels, allowing the human and robot to support each other in different ways, as needs and capabilities change throughout a task [1]. With the subsequent flexibility and adaptability of a human-robot collaboration team, production scenarios in permanently changing environments as well as the manufacturing of highly customized products become possible.

A. Motivation

To enable an effective collaboration, recent research in the field of psychology has focused on cognitive processes of joint-action among humans [2]. Psychological studies [3] show that in collaborating human teams, an effective coordination requires participants that plan and execute their actions in relation to what they anticipate from the other team member, and not just react on the other's current activities. Hence, for an efficient human-robot team, this knowledge needs to be transfered to a given set-up. The benefit of anticipatory action in a human-robot context is shown in [4], where a significant improvement of task efficiency compared to reactive behavior was possible. Following [5], a common representation of human and robot capabilities has to be found, because it is important for a collaborating team to know the skills of the other partner in order to assign certain difficult tasks according to specific skills correctly.

This work aims to integrate industrial robots in human-dominated working areas using multiple input modalities to allow a true peer-to-peer level of collaboration. Thus, a *smart working environment* for joint-action between a human and an industrial robot is presented as an experimental set-up consisting of various sensors monitoring the environment and the human worker, an industrial robot, an assembly line supplying the manufacturing process with material and tools, and a working table.

B. Related Work

Human-Robot collaboration using peer-to-peer approaches has lately become one of the central research issues in robotics.

[6] presents *KAMARA* a human-robot team using a multi-agent control architecture. Their mobile system was built up with a two-arm-manipulator and an overhead-camera. A proactive collaboration based on the recognition of intentions is described in [7]. Intention can be considered as a state of mind of the human that can not be measured directly. But vice versa, human action is a result of intention. Therefore [7] use Dynamic Bayesian Networks (DBNs) to deal with these uncertainties. [8] presents a cognitive architecture for a humanoid robot to allow it to interact with a human in a kitchen scenario. This architecture is organized in a hierarchical way, using three layers that specify the behavior of the robot in various situations. Leonardo [9] is a fully-embodied humanoid robot with social skills that enable the

978-1-4244-2213-5/08/$25.00 ©2008 IEEE

robot to learn and collaborate effectively in human settings. Mel the robotic penguin [10] acts as a host for a research lab, guiding visitors through the demonstration of a research prototype. The NASA peer-to-peer human-robot interaction system [11] is designed to allow humans and robots to collaborate on joint tasks: cooperation in this system mainly takes place when one agent asks another for help while dealing with a situation. The approach in [12] describes a scenario where a human and a robot system with two robotic arms build together a wooden model of an aircraft using an cognitive architecture divided into the high-level components *input, interpretation, representation, reasoning,* and *output* with several functional modules.

C. *Organization of the Paper*

The remainder of this paper is organized as follows: Section II describes in detail the system set-up and its high-level architecture. Section III gives an overview of the hardware and the used input modalities. The demonstration use case is specified in Section IV. Section V concludes the paper.

II. The Cognitive System

A. *Demonstration Scenario* - The Cognitive Factory

As defined in [13], cognitive systems follow longterm goals, e.g. to reach a destination by autonomous driving or the assembly of a product. Furthermore, a cognitive system perceives its environment via sensors, processes the sensor input and reacts to situations based on knowledge in an appropriate way. This behavior is achieved by closing the cognition loop *Perception, Cognition* and *Action* as depicted in Figure 1.

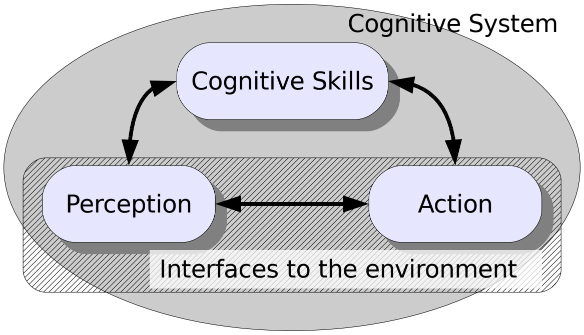

Fig. 1. Definition of cognitive systems using a closed loop of perception, cognition using cognitive skills, and action

One demonstrator of such cognitive systems in the *CoTeSys*-cluster is the *Cognitive Factory* [14] consisting of an automatic assembly station (project *CogMaSh - Cognitive Machine Shop*), purely human dominated assembly (project *ACIPE - Adaptive Cognitive Interaction in Production Environments* [15]) and hybrid assembly presented in this paper (project *JAHIR - Joint Action for Humans and Industrial Robots*). The benefits of all three assembly systems will be combined in later stages of expansion of the *Cognitive Factory* to support more effective manufacturing of a range of products.

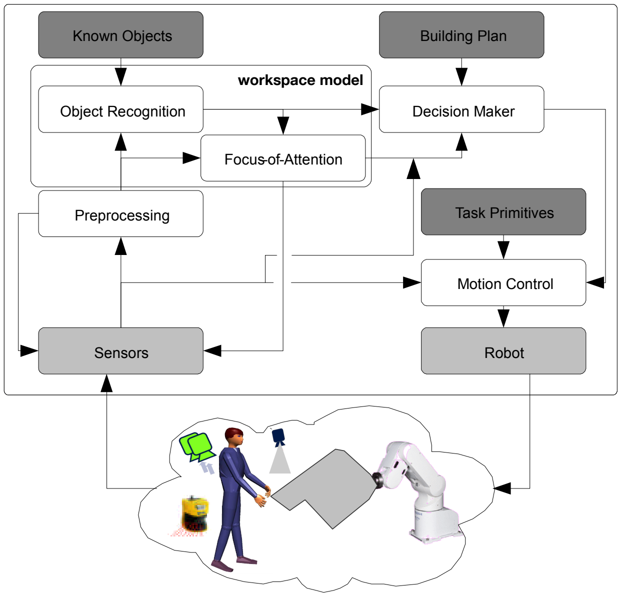

Fig. 2. Overview of the high-level architecture that enables joint action using multimodal observation cues. The light grey boxes represent the interfaces to the "real world". The dark gray boxes are the knowledge databases. The white boxes are the parts that enable together with the knowledge databases cognitive skills.

B. *Definition of the high-level architecture*

The high-level architecture was defined according to research results of cognitive neuro-scientists covering the aspects of successful joint action: joint attention, action observation, task sharing and action coordination [2]. In the following sections we will explain in detail the defined modules that can cover these aspects.

Figure 2 shows the modules of the defined JAHIR architecture and their connections among each other to realize joint action between human and robot.

1) Joint Attention: For cooperation between robot and human, it is important that both partners coincide on the same objects and topics and create a perceptual common ground and shared representation [2]. That means that the system has to be able to know where the human's focus-of-attention is. The focus-of-attention can be extracted from data gained by visual sensor devices (cameras) using tracking techniques (e.g. [16]) or from data gloves [17]. Therefore, the system needs to recognize e.g. a gesture as pointing gesture, compute the location indicated by the human worker and then transform this information to estimate his focus-of-attention. Thus, an internal system representation of the environment is needed that is always kept up to date with information from the sensors (*workspace model*).

In addition to pointing gestures, the head orientation of the human worker can be used to compute his focus-of-attention. According to behavioral studies [18], the head orientation is directly connected to the human's focus-of-attention. Getting the information about the head orientation can be done e.g. with tracking techniques using mutual information presented in [19].

The extraction will be done in the module *focus-of-attention* that loops back to the sensors to give directly control commands to a pan tilt unit to follow the human's focus-of-attention. Related work using attention has been done e.g. in [20], [21].

Joint attention implies that the both partners coincide on the same objects. In the current production scenario, these objects are the parts and tools that might be used. This knowledge is stored in the database *Known Objects*. To recognize objects a template based approach is used in the module *Object Recognition*.

2) Task Sharing: To manage a predefined goal, in this case the assembly of a product, the robot-system needs to know about the task in the production process as well as the human worker along with skill rated steps. Therefore, the representation of the tasks should be as generic as possible to be able to change the role allocation dynamically even during the production. The knowledge of the building steps, plans, and skill primitives is represented by the database module *Building Plan* which is part of the high-level architecture depicted in Figure 2.

3) Action Observation and Coordination: All information perceived by the *sensors* builds up an internal representation of the working environment, the *workspace model*. In addition to the perceived data, the *workspace model* may also carry information of the inventory (e.g. how much parts are stocked), the assembly line (e.g. which part is on the assembly line), and the worktable (e.g. information where parts that are already in use can be placed).

This information is used in the decision making process (*Decision Maker*) to decide the next action step along with the knowledge about the task from the *Building Plan* module, the sensor information, and the extracted features (e.g. the focus-of-attention). The system has to start the execution of the likeliest next step in order to reach a *anticipatory behavior* that is important for efficient collaboration. If new sensor-input changes the situation and thus the decision, the system needs to change its behavior seamlessly.

To control the movement of the robot (module *Motion Control*), real time control of the robot is needed. Because the robot and the human worker share the same work space, the movements of the robot cannot be programmed off-line for safety reasons. After starting to execute the next step twoards reaching the assembly goal, the robot has to be aware of obstacles (e.g. the human worker's hand, body, or arm) and react in such situations in real-time to avoid collisions with them. In Figure 2 this *reactive behavior* is represented by the direct connection of the *Sensors* module to the *Motion Control* of the robot.

C. System Architecture and Implementation

JAHIR uses a distributed system architecture divided into the modules presented in the preceding section with several submodules categorized in perceiving (sensors), cognition (knowledge database, processing units, and state-machine) and actuator modules (robot-control and User-Interfaces). Since the image-processing modules require a lot of computing power and controlling the robot on-line requires commands on a 7ms timing cycle, the system modules are divided among multiple computers. Communication between these modules is realized with the middleware *ICE - Internet Communication Engine* [22].

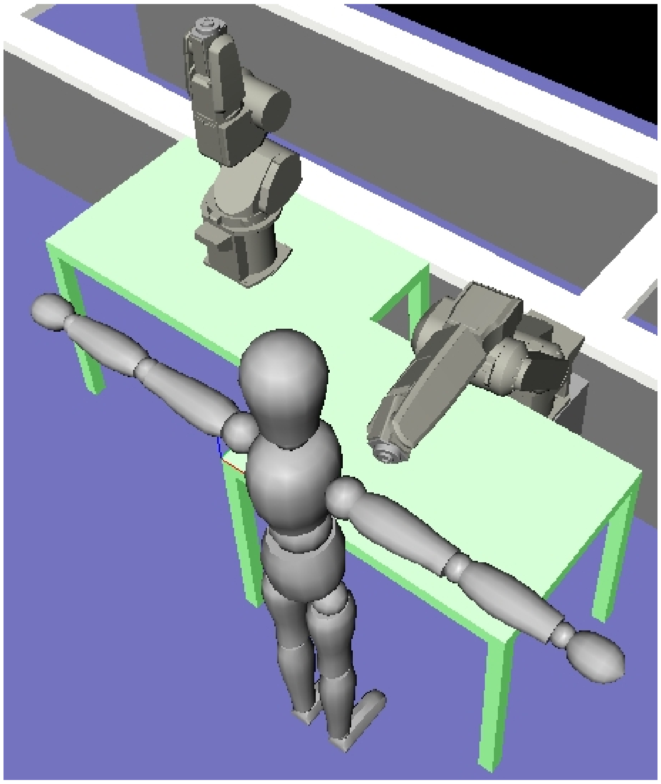

Fig. 3. Simulation of the experimental set-up

D. Knowledge and its Representation

As mentioned in Section II-B.2 the system needs to know about or be able to learn the assembly plan. It is also important to know what tools are available and what objects can be used on. This domain and task knowledge is represented in a first order logic in a knowledge database. Additional information perceived from the different sensors is also stored in the database for further reasoning purposes. For recognized containers on the workbench, information about their color, location, content and remaining number of parts are recorded as can be seen in this PROLOG-example code:

```
container(red, 44, 44, 'nuts', [12]).
container(blue, -9, -4, 'screws', [10]).
```

The system observes the humans activity within the shared workspace with multiple sensors and collects information about the current context. In combination with the knowledge about the next assembly steps, the system has a basis on which anticipation can be realised. An example is that the system has perceived that a certain part has been taken by the worker and knows that a tool is required for the next manufacturing step. The system can then try to locate this tool and, if successful, grab it and hand it over to the human. Combined with knowledge about the worker's current focus-of-attention, it might be necessary to attract his attention to the handover position. An example of retrieving what part has been grabbed by the user is given below:

```
gesture(grabbing, 42.72127, 45.55154).

taken(PART) :-
gesture(grabbing, XG, YG),
container(C, XC, YC, PART, [A]),
abs(XG-XC)<5,
abs(YG-YC)<5,
A>0,
A2 is A-1,
retract(container(C, XC, YC, PART, [A])),
assert(container(C, XC, YC, PART, [A2])),
assert(taken(PART)),
nl.

taken(’nuts’).

container(red, 44, 44, ’nuts’, [11]).
```

The sensors detect a gesture at the given example location *XG* and *YG*. The system is trying to find out if the user has taken a *PART* and what part he has taken. Therefore, the gesture has to be of type *grabbing* and needs to occur close enough to a container with color *C* at Position *XC* and *YC*. In addition, the number of parts *A* in this container has to be larger than zero. In the presented example the system then assumes that the part *nuts* has been picked and the remaining number of parts in this container is reduced to *A2*.

III. Shared Workbench for Joint Assembly

The set-up of the smart working environment depicted in the Figures 5 and 4consists of various sensors monitoring the human worker to anticipate and recognize his behavior and to be able to react in dangerous and unforeseen situations in an appropriate way. As assistant to the human worker in the production scenario, an industrial robot manipulator is used which has access to an assembly line and can nearly reach the whole workbench.

The industrial robot manipulator arm used in the set-up is a Mitsubishi robot RV-6SL. It can pick up objects with a maximum weight of six kilograms and a can move within a radius of 0.902 m. Furthermore, the robot is able to change the currently installed gripper by himself at a station to choose the one who fits best for the next task.

The following sections summarize the multimodal input modalities that can be extracted in the experimental set-up:

A. Visual Modality

In total four firewire cameras can be used in the smart working environment to gain information on the visual channel. One camera is mounted close to the tool center point of the manipulator and can be used for the detection of known objects and their position in the working environment to solve tasks including hand-over of parts and tools. To recognize objects lying on the desk the robot moves in a predefined position. The known objects are stored as templates in a database and are used in different scales and rotations to detect regions-of-interest and classify them using template-matching [23].

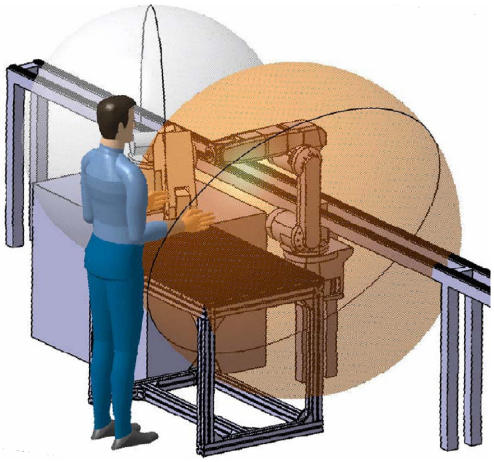

Fig. 4. Shared workspace of human and robot

As depicted in 5 the remaining three firewire-cameras are mounted on a cage that covers the workspace. Camera I and II are directed from left above and right above to the workspace. These calibrated cameras can be used for a non-invasive pose estimation and tracking of the human's hands in world-coordinates. The trajectories of the hands can be used to infer the human's intention as well as a safe motion planing for the robot. Camera III is directed from the front to record the human's head and estimate its position and rotation in space with mutual information [19] to estimate the focus-of-attention. Another possiblity using Camera III is the adaption of the system to a specific worker by recognizing specific faces. Examples for adaption are to offer a well trained worker less support than a newbie to the system, or to use a different hand-over position for a left-handed worker. The mounted cameras reveal the possibility to record the whole production process and use this data for off-line experiments, training, and the evaluation of adequate algorithms having always the same conditions.

B. Dataglove for Gestures

For detecting gestures like pointing or grasping, a P5 Dataglove is available which delivers the coordinates of the hand as well as information about finger bending. A disadvantage of this modality is that the accuracy of the position data is approximately 1 to 2 cm and the hand has to be closer than 1 meter to the sensor tower. Because a more reliable location of the hand position is required for the handover, an extra stereo infrared based marker tracking system with higher accuracy (0.1 cm) can be used. At the same time,

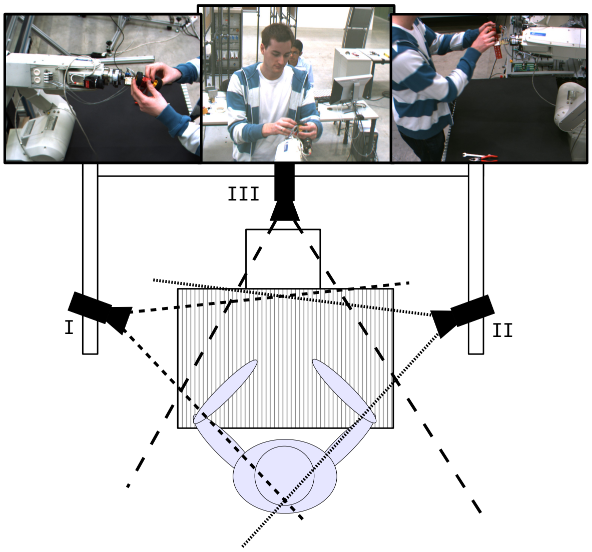

Fig. 5. Experimental set-up of the JAHIR smart working environment. Camera I and II are directed to the working table to track the hands of the human, camera III is directed to the head of the human worker to estimate its position and rotation.

this more robust tracking enables the observation of a larger workspace. However, due to working with infrared light the measurement is very sensitive to sunlight.

C. Photonic Mixer Device

For more complex image segmentation tasks, such as those needed for vision-based hand tracking or object recognition, the novel Photonic Mixer Device technology collects depth-information in real-time. The camera emits infrared light and measures the time-of-flight to calculate distances from the camera. It has a resolution of 64 x 48 Pixels at 25 frames per second. A more detailed description about this sensor and the used calibration techniques can be found in [24].

D. Speech Recognition

To enable control by voice, a speech recognition module will be integrated. Starting with a few commands to the system using a head-mounted microphone to prevent disturbances, results of another project (*MUDIS: A MUltimodal DIalogue System for Intuitive Human-Robot Interaction*) dealing with natural dialogues will be considered and integrated in the *JAHIR*-setup in the near future leading to a more natural and intuitive way of communication between user and robot.

E. Dealing with Data

The realtime-database presented in [25] will be used for realtime data recording. Currently, this database is used in cognitive vehicles to record large amounts of sensor data. Recording the perceived data brings up two major advantages:

- The recorded sensor-input can be taken for replay or simulation of certain situations. In addition, the gathered material can be analysed by humans e.g. to reveal gestures used in the production process, or to see if the worker is frightened.
- The data can be used for benchmarking purpose. Different implementations can be tested on the same data under the same conditions. After the new systems prove to work better than the old one, they can be used on the real set-up. The recorded database of sensor data can also be used by other projects to evaluate their system in an unknown environment.

IV. Use Case

The use case of the *JAHIR*-demonstrator will evolve in time as the capabilities of the system grow and more results of related projects will be integrated.

As initial demonstration scenario of joint-action, the assembly of parts of a LEGO Alpha-Rex has been chosen, inspired by the vision of *robots building robots*. Furthermore this product is easy to disassemble and therefore can be reused for experimental studies. It can also be adapted to the desired degree of complexity that seems reasonable. The focus of this use case is the hand-over of work-pieces and tools from a human to the robot and vice-versa. The assembly steps of the human worker can be seen in Figure 6.

In the beginning the robot picks the sensor and hands it over to the human. The same is done with the pre-assembled arm. The human either knows how to assemble the product or can retrieve the relevant assembly information from a display. However, he fetches screws and nuts from containers located on the table required to connect the sensor to the arm. The system monitors the human activities via the multimodal input channels listed in Section III. The system realizes that a wrench could be useful for the next workstep. The tool is located by the system and handed over to the worker. The human fixes the sensor to the arm. Afterwards, the Alpha-Rex-Body is handed over from the robot to the human. The arm is mounted to the body and the sensor is connected with a cable to the core of the Alpha-Rex. The finished Alpha-Rex is handed over to the robot.

V. Conclusions

We have presented the design of a joint-action assembly demonstrator that allows cooperation between humans and robots in a shared workspace. The main idea of the system is to enable peer-to-peer collaboration using multiple input modalities and a cognitive backbone. Efficiency in the collaboration can be reached with an anticipatory behaviour of the robot system which is combined with reactive behaviour to guarantee safety.

VI. Acknowledgement

This ongoing work is partly supported by the DFG excellence initiative research cluster *Cognition for Technical Systems CoTeSys*, see www.cotesys.org for further details.

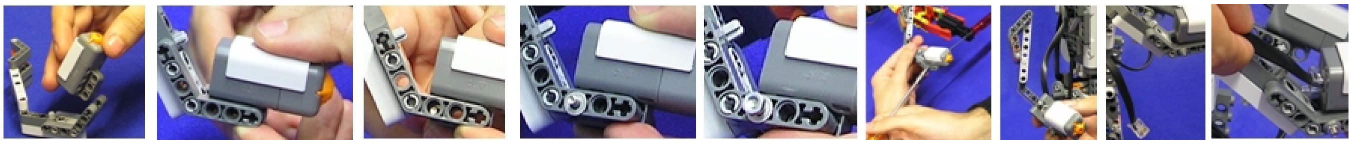

Fig. 6. Steps of assembling an Alpha-Rex-Arm

References

[1] J. L. Marble, D. J. Bruemmer, D. A. Few, and D. D. Dudenhoeffer, "Evaluation of supervisory vs. peer-peer interaction with human-robot teams," in *HICSS '04: Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 5*. Washington, DC, USA: IEEE Computer Society, 2004, p. 50130.2.

[2] N. Sebanz, H. Bekkering, and G. Knoblich, "Joint action: bodies and minds moving together," *Trends in Cognitive Sciences*, vol. 10, no. 2, pp. 70–76, February 2006.

[3] G. Knoblich and J. S. Jordan, "Action coordination in groups and individuals: learning anticipatory control." *J Exp Psychol Learn Mem Cogn*, vol. 29, no. 5, pp. 1006–1016, September 2003.

[4] G. Hoffman and C. Breazeal, "Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team," in *HRI '07: Proceeding of the ACM/IEEE international conference on Human-robot interaction*. New York, NY, USA: ACM, 2007, pp. 1–8.

[5] F. Tang and L. E. Parker, "Peer-to-peer human-robot teaming through reconfigurable schemas," in *AAAI Spring Symposium on "To Boldly Go Where No Human-Robot Team Has Gone Before"*, Stanford University, March 2006.

[6] T. Laengle, T. Hoeniger, and L. Zhu, "Cooperation in human-robot-teams," in *Proceedings of the IEEE International Symposium on Industrial Electronics, 1997. ISIE '97.*, 7-11 July 1997, pp. 1297–1301vol.3.

[7] O. Schrempf, U. Hanebeck, A. Schmid, and H. Worn, "A novel approach to proactive human-robot cooperation," in *IEEE International Workshop on Robot and Human Interactive Communication, 2005. ROMAN 2005.*, 13-15 Aug. 2005, pp. 555–560.

[8] C. Burghart, R. Mikut, R. Stiefelhagen, T. Asfour, H. Holzapfel, P. Steinhaus, and R. Dillmann, "A cognitive architecture for a humanoid robot: a first approach," in *Humanoid Robots, 2005 5th IEEE-RAS International Conference on*, 5-7 Dec. 2005, pp. 357–362.

[9] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Chilongo, "Tutelage and collaboration for humanoid robots," *International Journal of Humanoid Robotics*, vol. 1, no. 2, pp. 315–348, 2004.

[10] C. L. Sidner and M. Dzikovska, "A first experiment in engagement for human-robot interaction in hosting activities," in *Advances in Natural Multimodal Dialogue Systems*, N. Bernsen, L. Dybkjær, and J. van Kuppevelt, Eds. Springer, 2005.

[11] T. W. Fong, C. Kunz, L. Hiatt, and M. Bugajska, "The human-robot interaction operating system," in *Proceedings of the International Conference on Human-Robot Interaction*. ACM, 2006.

[12] M. Rickert, M. Foster, M. Giuliani, T. By, G. Panin, and A. Knoll, "Integrating language, vision and action for human robot dialog systems," in *Proceedings of the International Conference on Human-Computer Interaction*, C. Stephanidis, Ed. Beijing: Springer, July 2007, pp. 987–995.

[13] M. Buss, M. Beetz, and D. Wollherr, "Cotesys - cognition for technical systems," in *Proceedings of the 4th COE Workshop on Human Adaptive Mechatronics (HAM)*, 2007.

[14] M. Zäh, C. Lau, M. Wiesbeck, M. Ostgathe, and W. Vogl, "Towards the Cognitive Factory," in *International Conference on Changeable, Agile, Reconfigurable and Virtual Production (CARV)*, Toronto, Canada, July 2007.

[15] M. F. Zäh, M. Wiesbeck, F. Engstler, F. Friesdorf, A. Schubö, S. Stork, A. Bannat, and F. Wallhoff, "Kognitive Assistenzsysteme in der Manuellen Montage," in *wt Werkstattstechnik online*, vol. 97, 9. Springer-VDI-Verlag, 2007, pp. 644–650.

[16] A. A. Argyros and M. I. A. Lourakis, "Real-time tracking of multiple skin-colored objects with a possibly moving camera," in *ECCV (3)*, 2004, pp. 368–379.

[17] S. Reifinger, F. Wallhoff, M. Ablaßmeier, T. Poitschke, and G. Rigoll, "Static and dynamic hand-gesture recognition for augmented reality applications," in *Proceedings of the International Conference on Human-Computer Interaction*, C. Stephanidis, Ed. Beijing: Springer, July 2007.

[18] N. J. Emery, "The eyes have it: the neuroethology, function and evolution of social gaze," pp. 581–604, August 2000. [Online]. Available: http://www.ingentaconnect.com/content/els/01497634/2000/00000024/00000006/art00025

[19] G. Panin and A. Knoll, "Mutual information-based 3d object tracking," *International Journal of Computer Vision (IJCV)*, 2007.

[20] G. Heidemann, R. Rae, H. Bekel, I. Bax, and H. Ritter, "Integrating context-free and context-dependent attentional mechanisms for gestural object reference," *Mach. Vision Appl.*, vol. 16, no. 1, pp. 64–73, 2004.

[21] A. Edsinger, "Robot manipulation in human environments," Ph.D. dissertation, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2007.

[22] Zeroc, "Internet communications engine," http://www.zeroc.com.

[23] T. Müller, P. Ziaie, and A. Knoll, "A wait-free realtime system for optimal distribution of vision tasks on multicore architectures," in *Proc. 5th International Conference on Informatics in Control, Automation and Robotics*, 2008.

[24] F. Wallhoff, M. Ruß, G. Rigoll, J. Göbel, and H. Diehl, "Surveillance and activity recognition with depth information," in *IEEE International Conference on Image Processing (ICIP)*, San Antonio, Texas, USA, September, 16-19 2007.

[25] M. Goebl and G. Färber, "A real-time-capable hard- and software architecture for joint image and knowledge processing in cognitive automobiles," in *Intelligent Vehicles Symposium*. IEEE Press, June 2007, pp. 737–740.