Combining Goal Inference and Natural-Language Dialogue for Human-Robot Joint Action

Mary Ellen Foster[1] and **Manuel Giuliani**[1] and **Thomas Müller**[1] and **Markus Rickert**[1] and **Alois Knoll**[1]
Wolfram Erlhagen[2] and **Estela Bicho**[2] and **Nzoji Hipólito**[2] and **Luis Louro**[2]

Abstract. We demonstrate how combining the reasoning components from two existing systems designed for human-robot joint action produces an integrated system with greater capabilities than either of the individual systems. One of the systems supports primarily non-verbal interaction and uses dynamic neural fields to infer the user's goals and to suggest appropriate system responses; the other emphasises natural-language interaction and uses a dialogue manager to process user input and select appropriate system responses. Combining these two methods of reasoning results in a robot that is able to coordinate its actions with those of the user while employing a wide range of verbal and non-verbal communicative actions.

1 INTRODUCTION AND MOTIVATION

As robot systems become increasingly sophisticated, their role is moving from one where the robot is essentially an intelligent tool to one where the robot is able to participate as a full team member in collaborative tasks. Supporting this type of human-robot cooperation requires that the robot system be able to produce and understand a wide range of natural communicative cues in order to allow humans to cooperate with it easily. For example, [15] experimentally demonstrated the contribution of anticipatory action to the fluency of human-robot interaction; similarly, natural-language dialogue has been shown to be an effective means of coordinating actions between a human and a robot [7].

A number of previous systems have also addressed the task of human-robot cooperation, using a variety of communicative styles. The Leonardo robot [2], for example, is able to learn simple action sequences and to execute them jointly with the user. The Ripley system [24] is able to manipulate objects in response to spoken requests from a human partner; a more recent robot from the same group [16] increases the responsiveness of the system and allows the action planner to adapt flexibly to a rapidly-changing world. The BARTHOC [26] and ARMAR [27] humanoid robots both support multimodal dialogue to interact with a human user in a variety of settings and domains. The experiments described in [15] demonstrated that understanding and anticipating the user's actions produces a robot that can cooperate more smoothly with a human user.

Since an intelligent robot system must both process continuous sensor data and reason about discrete concepts such as plans, actions, and dialogue moves, this type of system is often made up of components drawing from an assortment of representation and reasoning paradigms. The robot system described in [22], for example, combines low-level robot control and vision systems with a high-level planner, using connectionist kernel perceptron learning to learn the effects of different domain actions. Integration among the different components of this system is achieved through a common representation of actions and their effects. Such hybrid architectures are also particularly common when the robot is designed to cooperate with a human partner; recent examples include [13, 17, 32].

In this paper, we present two robot systems designed to cooperate with humans on mutual tasks and then show how combining reasoning components from these systems results in a more powerful integrated system. Both of the robot systems have been developed in the context of the JAST[3] project ("Joint Action Science and Technology"). The two main goals of this project are to investigate the cognitive, neural, and communicative aspects of jointly-acting agents, both human and artificial, and to build jointly-acting autonomous systems that communicate and work intelligently on mutual tasks. The common task across the project is *joint construction*—that is, multiple agents working together to assemble objects from their components.

The two JAST human-robot systems support intelligent cooperation with humans on this joint construction task. Although both systems address the same basic task and incorporate similar input- and output-processing components, the reasoning components are implemented using very different techniques and they support very different styles of interaction. The *goal inference* system is implemented using dynamic neural fields and concentrates on inferring the user's intended domain actions based on their non-verbal behaviours and on selecting appropriate domain actions for the system to perform in response. For example, if the user picks up a bolt in a way that indicates that they intend to use it themselves, the system might pick up the corresponding wheel and hold it out to the user. The *dialogue* system, on the other hand, concentrates on understanding and generating multimodal natural-language utterances to support cooperation between the human and the robot, using a dialogue manager. Sections 2–3 present the details of these two systems and show a typical interaction with each.

Since the two JAST human-robot systems address the same task, using complementary forms of reasoning, it is possible to combine the two forms of reasoning into a single system. This integrated system is able both to intelligently infer the user's actions and suggest appropriate responses, and also to engage in dialogue with the user to support coordination and to discuss situations when the system is unable to infer the user's goal. In Section 4, we present this integrated system and show a sample of the interactions that it can support that are not possible with either of the individual systems; this section also gives some technical details of how the components of the two

[1] Technische Universität München, Germany, contact: foster@in.tum.de
[2] University of Minho, Portugal, contact: wolfram.erlhagen@mct.uminho.pt

[3] http://www.euprojects-jast.net/

systems are combined in practice. Finally, in Section 5, we compare the integrated system with other similar systems and summarise the contributions of the system and the areas for future work.

2 GOAL INFERENCE BASED ON DYNAMIC NEURAL FIELDS

The first of the JAST human-robot systems concentrates on giving the robot the ability to predict the consequences of observed actions, using an implementation inspired by neurocognitive mechanisms underlying this capacity in humans and other social species. Many contemporary theories of intention understanding in familiar tasks rely on the notion that an observer uses their own motor repertoire to simulate an observed action and its effect ([4], for a review see [25]). The selection of an appropriate complementary behaviour in a joint action task depends not only on the inferred goal of the partner, but also on the integration of additional information sources such as shared task knowledge (e.g., a construction plan) and contextual cues.

The cognitive control architecture for action coordination in the joint construction scenario is formalized by a coupled system of dynamic fields representing a distributed network of local but connected neural populations [3]. Different pools of neurons encode task-relevant information about action means, action goals, and context in the form of activation patterns that are self-sustained through recurrent interactions.

The motor simulation idea is implemented by the propagation of activity through interconnected neural populations that constitute a learned chain of motor primitives directed towards a specific goal [6]. Typical examples in the context of the construction scenario are reaching-grasping-placing/plugging sequences. The chains are automatically triggered by an observed motor act (e.g., reaching or grasping) whenever additional input from connected dynamic fields (e.g., representing the currently available subgoals) pre-activates the neural populations. As a consequence of the motor simulation, the robot is able to react to the partner's action sequences well ahead of their completion. This anticipation capacity has been shown to be crucial for a fluent team performance [1, 15].

In the layer of the control architecture linked to motor execution, neural populations represent the decision about the most appropriate complementary behaviour. The behaviour is selected as a consequence of a competition process between all response alternatives getting input from connected layers (for details see [1]).

A system based on this dynamic field architecture was implemented to support human-robot cooperation on the JAST joint construction task. This system constructs a toy vehicle (Figure 1) with the user. The vehicle is composed of several components which are initially distributed in the separated working areas of the two teammates; this ensures that neither of the agents is able to reach all of the required components on its own and must rely on the partner to retrieve them, making joint action essential to a successful interaction. The robotics platform we are currently using consists of a torus on which are mounted a 7 DOFs AMTEC arm (Schunk GmbH & Co.KG) with a 3-fingered BARRET hand (Barrett Technology Inc.) and a stereo vision system. The system uses synthesised speech to communicate its reasoning process to the human partner.

To control the arm-hand system, we applied a global planning method in posture space that facilitates the integration of optimization principles derived from experiments with humans [5]. For the object recognition as well as for the classification of object-directed hand postures and communicative gestures such as pointing or demanding an object, a combination of feature- and correspondence-based pattern recognizers were used [30]. As a software development platform we haven chosen YARP [19]. This open-source project supports inter-process communication, image processing and a class hierarchy to ease code reuse across different hardware platforms.

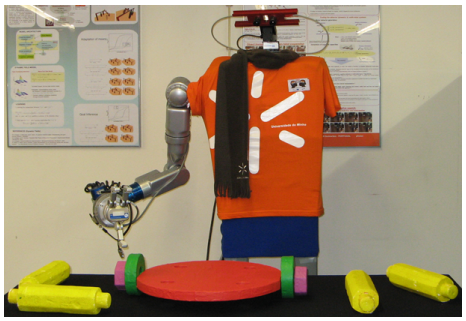


Figure 1. The JAST goal-inference robot together with the toy vehicle that the human and the robot jointly construct.

Figure 2 illustrates a typical example of the goal inference and action selection capacities in this domain. In the top image, the human reaches his open hand towards the robot teammate. By activating the respective action chain in its repertoire, the robot interprets this gesture as a request for a bolt to fix the wheel. Since the human has already mounted the wheel on his side of the construction, this inferred goal describes a currently active subtask. A logical complementary action sequence would be that the robot grasps a bolt to place it in the teammate's hand. However, the human has seemingly overlooked a bolt in his own working area. In this situation, the representation of the inferred goal together with the representation of the bolt in the work space of the human trigger the decision to make a pointing gesture directed towards the object. In addition, the robot uses speech to explain the type of error the human is making.

3 DIALOGUE-BASED HUMAN-ROBOT INTERACTION

Like the system described in the preceding section, the JAST human-robot dialogue system [23] also supports multimodal human-robot collaboration on a joint construction task. In this case, the user and the robot work together to assemble wooden construction toys on a common workspace, coordinating their actions through speech, gestures, and facial displays. The robot (Figure 3) consists of a pair of Mitsubishi manipulator arms with grippers, mounted in a position to resemble human arms, and an animatronic talking head [29] capable of producing facial expressions, rigid head motion, and lip-synchronised synthesised speech. The input channels consist of speech recognition, object recognition, robot sensors, and face tracking; the outputs include synthesised speech, head motions, and robot actions. The components of the system communicate with each other using the Ice distributed object middleware system [14].

The robot is able to manipulate objects in the workspace and to perform simple assembly tasks. The primary form of interaction with the current version of the system is one in which the robot instructs the user on building a particular compound object, explaining the necessary assembly steps and retrieving pieces as required, with the user performing the actual assembly actions. As with the dynamic-field system, the workspace is divided into two areas—one belonging to the robot and one to the human—in order to make joint action necessary for success in the overall task.

Input on each of the channels is processed using a dedicated module for that channel. To process the speech, we use a Java Speech

Figure 2. Example of the goal inference (top) and action selection (bottom) capacities which are implemented by the dynamic field architecture. The robot uses speech to communicate the results of its reasoning about the behaviour of the teammate.

Figure 3. The JAST dialogue robot with a selection of wooden construction-toy components.

API interface to the commercial Dragon NaturallySpeaking speech recogniser [21]. A camera mounted above the common work area provides two-dimensional images of the contents of the workspace. The information from this camera is pre-processed to extract regions of interest (ROIs). The extracted ROIs are then processed in parallel by a template-based object-recognition module [20] and a module that performs static hand-gesture recognition [33].

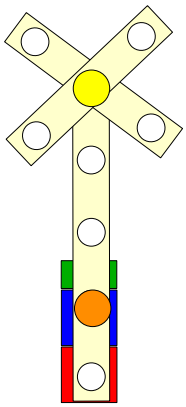
Input received on all of the input sensors is continuously processed by the corresponding modules and broadcast through Ice, using the built-in *IceStorm* publish-subscribe mechanism. All of the input messages are received by a multimodal fusion component [11, 12], which parses the recognized speech into logical forms using the OpenCCG grammar formalism [31] and combines it with the recognised non-verbal behaviour to produce multimodal hypotheses representing user requests. The fusion hypotheses are then sent to the dialogue manager, which selects an appropriate response.

The dialogue manager is implemented using the TrindiKit dialogue management toolkit [18]. This toolkit uses the well-known *information-state update* approach to dialogue management [28], which allows a dialogue to be modelled declaratively. When the dialogue manager receives a new set of fusion hypotheses, it selects the appropriate system response using information from three sources: the inventory of objects in the world, a representation of the current assembly state, and the history of the dialogue. When the system is jointly following an assembly plan with the user, the dialogue manager is able to select from different strategies for traversing the plan: it may use a *postorder* strategy, in which it proceeds directly to describing the concrete assembly actions, or it may use a *preorder* strategy, in which the structure of the plan is described before giving specific assembly actions. More details on the dialogue manager and on the description strategies are given in [10].

Once the dialogue manager has selected a response to the user's multimodal utterance, it sends the specification of the response to the output planner. This module in turn sends commands to select appropriate output on each of the individual channels to meet the specification: linguistic content including appropriate multimodal referring expressions [9], facial expressions and gaze behaviours of the talking head [8], and actions of the robot manipulators. The user then responds to the system utterance by speaking or performing actions in the world, and the interaction continues until the target object has been assembled.

An excerpt from a typical interaction between a user and the JAST dialogue system is shown in Figure 4. In this excerpt, the robot knows the full plan for building the target object: a "railway signal", which has sub-components called a "snowman" and a "flower". The assembled object is shown in Figure 4(a). In the excerpt, the robot instructs the user on how to build the target object, using a preorder strategy, and the user learns to make particular sub-components along the way. We are currently carrying out a system evaluation based on this robot-as-instructor scenario. The evaluation is designed to compare the two description strategies in terms both of user satisfaction and in success in the overall joint-construction task.

We will shortly extend the dialogue system to handle scenarios where the user also knows the assembly plan. In such situations, the main goal of the interaction is no longer instruction, but rather—as with the goal-inference system described previously—coordination between the partners, and the user will be able to take much more initiative in the dialogue than is currently possible. We will return to the details of this extended scenario in the following section.



(a) "Railway signal"

- SYSTEM: To build a railway signal, we need to make a snowman and a flower.
- USER: How do you make a snowman?
- SYSTEM: *[Picking up a red cube and holding it out]* To make a snowman, insert the green bolt into your blue cube and fasten it with this cube.
- USER: *[picks up bolt and cube; takes other cube; assembles "snowman"]* Okay.
- SYSTEM: Well done, you have made a snowman. Now we need to make a flower. To make a flower, [. . .]

(b) Dialogue excerpt

Figure 4. A sample object and an excerpt from an interaction where the robot instructs the user on how to construct this object.

4 INTEGRATING GOAL INFERENCE AND NATURAL-LANGUAGE DIALOGUE

There are a number of similarities between the two human-robot systems described above. Both support the same basic task—joint construction—and view the goals and subgoals of this task in a similar way. Also, the input and output channels used by the two systems are very similar: both include object and gesture recognition in the input and produce speech and robot-manipulator actions as part of their output. On the other hand, the reasoning processes used by the two systems are very different: the former uses dynamic neural fields to perform goal inference and action selection based entirely on non-verbal input, while the latter uses techniques from issue-based dialogue management to engage in natural-language conversation with some multimodal components. The strengths of the two systems are also complementary: the dynamic-field system is good at detecting and reasoning about the user's non-verbal actions, but uses language only for a limited form of canned output; the dialogue system supports advanced linguistic interaction, but has no mechanism to infer the user's intention from their actions in the world.

Motivated by the above similarities and complementary features, we have combined components from the two individual human-robot systems into a single, integrated architecture. The hardware platform for the integrated system is the robot from the dialogue system (Figure 3), while the scenario is an extended version of the scenarios used by each of the individual systems. As in the dynamic-field scenario, the user and the robot are both assumed to know the assembly plan for the target object and are able to infer the partner's intentions based on their behaviour, and the main goal of the interaction is for the two participants to coordinate their actions. As in the dialogue system, this coordination is accomplished through natural-language dialogue incorporating both verbal and non-verbal communication.

Figure 5 shows the high-level architecture of the integrated system. Messages on all of the multimodal input channels (speech, gestures, and recognised objects) are sent to both of the input-processing components, each of which—just as in the individual systems—reasons about the meaning of the user's actions in the current context, each drawing information from the same set of state modules (plan state, object inventory, interaction history). The inferred goals and suggested system responses from the goal-inference system are then passed to the the dialogue manager, which incorporates this information along with the processed messages from the fusion system into the (extended) information state of the integrated system. The dialogue manager then uses enhanced update rules to select an appropriate system response to the input. Finally, just as in the individual systems, the selected response is sent to the output system for realisation on the output channels.

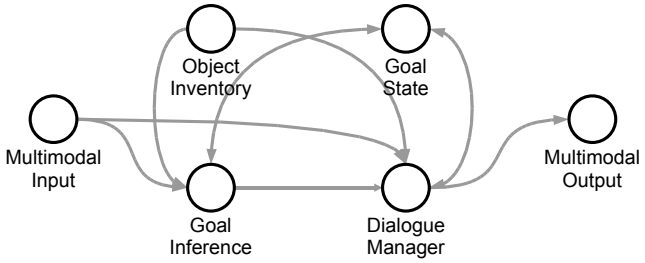


Figure 5. The architecture of the integrated system.

This integrated system supports interaction patterns that would not be possible with either of the individual systems. Most importantly, it is able both to detect unexpected actions from the user (i.e., actions that do not meet what it believes to be the current subgoals) and to engage the user in dialogue to discuss how to deal with the unexpected action. When both forms of reasoning work together, the system is able to detect such user actions and to produce a variety of responses, including correcting the user, asking for clarification as to the user's intentions, or attempting to silently revise its representation of the goal state. Varying the system's response to this situation is able to produce systems with different interactive "personalities", ranging from one that always makes the user follow the plan selected by the system to one where the user has full control over the course of the interaction.

Figure 6 shows a sample interaction between a user and the integrated system, where the role of each of the reasoning components is shown throughout. In this interaction, the user and the robot are jointly building the "railway signal" object (Figure 4(a)). At the start, the robot system has assumed that the user is building the "snowman" sub-component. When the user grasps a medium slat, which is not needed for that subgoal, the goal inference system detects this (just as in the sample interaction described at the end of Section 2) and sends a message to the dialogue manager that the user's action cannot be integrated into the current plan.

At this point, the system has several options to deal with the mismatch between its beliefs about the current subgoals and the recent action of the user. It might silently revise its view of the current subgoals, for example, or it might—as in Figure 2—correct the user's apparent "error". In the example, the system uses a third strategy, and one that is only available because of the integration of the dialogue components: it asks the user in words to clarify their intentions. After the user provides the needed clarification, also verbally, the dialogue manager updates the system's subgoals and informs the goal-inference system of the change. The goal-inference system then anticipates that, to meet this new subgoal, the user will need the nut that is lying on the robot's side of the table. The system therefore picks up the nut and offers it to the user without being asked.

As can be seen by the right-hand columns in Figure 6, this type of interaction would not be possible with either of the individual systems. The dialogue system does not have the necessary mechanism to infer the user's goals from their actions, while the goal-inference system would only have been able to respond to the user's unexpected

Actions	Dialogue Manager	Goal Inference
User grasps a medium slat		
		Notices that action does not meet current subgoal
	Tells output planner to ask for clarification	
SYSTEM: "We don't need a medium slat for the snowman"		
USER: "Yes, but I want to build the flower now"		
	Interprets response and updates subgoals	
		Suggests system response
	Sends message to output planner	
Robot picks up a nut and holds it out SYSTEM: "Then you'll need this nut"		

Figure 6. A sample interaction with the integrated system, showing the role of each individual reasoning component in the decision-making process.

action by treating it as an error rather than discussing the user's goals as in the example. Only when these two components are combined is this rich interaction made possible.

4.1 Technical Details

The two individual systems use the same basic information in their reasoning (task goals and subgoals, object inventory, input events); however, due to the different implementations, they represent this information quite differently. Also, at the implementation level, the components of the dynamic-field system use YARP to communicate with one another, while the dialogue system uses Ice as an integration platform. A specific goal of the integration has been to make as few changes as possible to the individual systems. An important aspect of creating the integrated system has therefore been coming up with a common representation for all of the relevant information, where the representation is compatible with both of the systems and both of the integration platforms.

To support the integration, we have defined generic interfaces to represent recognised gestures and objects, as well as inferred and proposed domain actions. These representations include the following information:

- The **Gestures** representation includes the type of gesture recognised (pointing, grasping, holding-out, unknown) and if necessary, the object indicated.
- The **Objects** representation includes the classification of the object, a 3D position and a flag indicating whether the object can be reached by the robot.
- The **Action** representation consists of the type of action (grasp-and-give, demand-and-receive, speak, undefined) and a string containing further specifications (e.g. the object-id for grasp-and-give or the sentence to speak out loud).

Internal communication between YARP and Ice is implemented via a connector module that translates Ice messages to YARP messages and vice versa.

5 DISCUSSION

We have presented two human-robot systems, each of which is designed to support the same joint construction task. One system uses dynamic neural fields to perform non-verbal goal inference and action selection, while the other uses a dialogue manager to support multimodal natural-language interaction. We have then shown how a system integrating the reasoning components of the two individual systems is able to take advantage of the complementary strengths of each to support interactions that neither system is able to support on its own. In particular, this integrated system is able both to detect the user's intentions and anticipate their needs, and to use natural-language dialogue to manage the joint activity. The integration of these two systems is made possible through well-defined interfaces that allow the two sets of reasoning components to share information about world state, task goals, and input events.

In contrast to the other systems mentioned in the introduction, the integrated JAST system is unique in that it combines methods and techniques taken from two separate, fully-implemented, existing systems—a neuro-inspired perception-action system and a symbolic, multimodal-dialogue system—to produce an integrated robot system that is able both to communicate with its human partner using language and to intelligently understand and anticipate the partner's intentions. As demonstrated by the example in the preceding section, the integrated system is able to go beyond the capabilities of either of the individual systems to support intelligent human-robot cooperation on the joint construction task.

The integrated system is currently under development: the necessary interfaces have been specified as described in Section 4.1, and the reasoning modules from the two systems are being adapted to use the common interfaces. When the is completed, we will run a user evaluation of the full system similar to that currently under way for the dialogue system to demonstrate the contribution of both forms of reasoning to natural human-robot joint action.

ACKNOWLEDGEMENTS

This work was supported by the EU FP6 IST Cognitive Systems Integrated Project "JAST" (FP6-003747-IP), `http://www.euprojects-jast.net/`. Thanks to the CIMA workshop reviewers for their useful comments and suggestions.

REFERENCES

[1] E. Bicho, L. Louro, N. Hipólito, and W. Erlhagen, 'A dynamic neural field architecture for flexible and fluent human-robot interaction.', in *Proceedings of the 2008 International Conference on Cognitive Systems*, pp. 179–185, (2008).

[2] 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*, **1**(2), 315–348, (2004).

[3] W. Erlhagen and E. Bicho, 'The dynamic neural field approach to cognitive robotics', *Journal of Neural Enginering*, **3**, R36–R54, (2006).

[4] W. Erlhagen, A. Mukovskiy, and E. Bicho, 'A dynamic model for action understanding and goal-directed imitation.', *Brain Research*, **1083**, 174–188, (2006).

[5] W. Erlhagen, A. Mukovskiy, E. Bicho, G. Panin, C. Kiss, A. Knoll, H. van Schie, and H. Bekkering, 'Goal-directed imitation for robots: a bio-inspired approach to action understanding and skill learning', *Robotics and Autonomous Systems*, **54**, 353–360, (2006).

[6] W. Erlhagen, A. Mukovskiy, F. Chersi, and E. Bicho, 'On the development of intention understanding for joint action tasks', in *Proceedings of the 6th IEEE International Conference on Development and Learning*, (July 2007).

[7] T. Fong, C. Thorpe, and C. Baur, 'Collaboration, dialogue, and human-robot interaction', in *Robotics Research*, volume 6 of *Springer Tracts in Advanced Robotics*, 255–266, Springer, (2003).

[8] M. E. Foster, 'Roles of a talking head in a cooperative human-robot dialogue system', in *Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA 2007)*, (September 2007).

[9] M. E. Foster, E. G. Bard, R. L. Hill, M. Guhe, J. Oberlander, and A. Knoll, 'The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue', in *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction (HRI 2008)*, (March 2008).

[10] M. E. Foster and C. Matheson, 'Following assembly plans in cooperative, task-based human-robot dialogue', in *Proceedings of Londial 2008*, (June 2008).

[11] M. Giuliani and A. Knoll, 'Integrating multimodal cues using grammar based models', in *Proceedings of HCI International 2007*, (July 2007).

[12] M. Giuliani and A. Knoll. MultiML – A general purpose representation language for multimodal human utterances, 2008. Submitted.

[13] N. Hawes, A. Sloman, J. Wyatt, M. Zillich, H. Jacobsson, G.-J. Kruijff, M. Brenner, G. Berginc, and D. Skočaj, 'Towards an integrated robot with multiple cognitive functions', in *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI 2007)*, (2007).

[14] M. Henning, 'A new approach to object-oriented middleware', *IEEE Internet Computing*, **8**(1), 66–75, (Jan–Feb 2004).

[15] G. Hoffman and C. Breazeal, 'Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team', in *Proceedings of the 2nd ACM/IEEE International Conference on Human Robot Interaction (HRI 2007)*, (2007).

[16] K.-y. Hsiao, S. Vosoughi, S. Tellex, R. Kubat, and D. Roy, 'Object schemas for responsive robotic language use', in *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction (HRI 2008)*, (2008).

[17] W. G. Kennedy, M. D. Bugajska, M. Marge, W. Adams, B. R. Fransen, D. Perzanowski, A. C. Schultz, and J. G. Trafton, 'Spatial representation and reasoning for human-robot collaboration', in *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI 2007)*, (2007).

[18] S. Larsson and D. R. Traum, 'Information state and dialogue management in the TRINDI dialogue move engine toolkit', *Natural Language Engineering*, **6**, 323–340, (2000).

[19] G. Metta, P. Fitzpatrick, and L. Natale, 'YARP: Yet another robot platform', *International Journal of Advanced Robotics Systems*, **3**(1), 43–48, (2006).

[20] T. Müller, P. Ziaie, and A. Knoll, 'A wait-free realtime system for optimal distribution of vision tasks on multicore architectures', in *Proceedings of the 5th International Conference on Informatics in Control, Automation and Robotics*, (May 2008).

[21] Nuance Communications. Dragon NaturallySpeaking 9. `http://www.nuance.com/naturallyspeaking/`.

[22] R. Petrick, D. Kraft, K. Mourão, C. Geib, N. Pugeault, N. Krüger, and M. Steedman, 'Representation and integration: Combining robot control, high-level planning, and action learning', in *Proceedings of the International Cognitive Robotics Workshop (CogRob 2008) at ECAI 2008*, (July 2008).

[23] M. Rickert, M. E. Foster, M. Giuliani, T. By, G. Panin, and A. Knoll, 'Integrating language, vision and action for human robot dialog systems', in *Proceedings of HCI International 2007*, (July 2007).

[24] D. Roy, K.-Y. Hsiao, and N. Mavridis, 'Mental imagery for a conversational robot', *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, **34**(3), 1374–1383, (June 2004).

[25] N. Sebanz, H. Bekkering, and G. Knoblich, 'Joint action: bodies and minds moving together', *Trends in Cognitive Sciences*, **10**, 70–76, (2006).

[26] T. Spexard, M. Hanheide, and G. Sagerer, 'Human-oriented interaction with an anthropomorphic robot', *IEEE Transactions on Robotics*, **23**(5), 852–862, (October 2007).

[27] R. Stiefelhagen, H. Ekenel, C. Fugen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A. Waibel, 'Enabling multimodal human-robot interaction for the Karlsruhe humanoid robot', *IEEE Transactions on Robotics*, **23**(5), 840–851, (October 2007).

[28] D. Traum and S. Larsson, 'The information state approach to dialogue management', in *Current and New Directions in Discourse and Dialogue*, eds., J. C. J. Van Kuppevelt and R. W. Smith, 325–353, Kluwer Academic Publishers, (2003).

[29] A. J. N. van Breemen, 'iCat: Experimenting with animabotics', in *Proceedings of the AISB 2005 Creative Robotics Symposium*, (2005).

[30] G. Westphal, C. von der Malsburg, and R. Würtz, 'Feature-driven emergence of model graphs for object recognition and categorization', in *Applied Pattern Recognition*, eds., A. Kandel, H. Bunke, and M. Last, 155–199, Springer Verlag, (2008).

[31] M. White, 'Efficient realization of coordinate structures in Combinatory Categorial Grammar', *Research on Language and Computation*, **4**(1), 39–75, (2006).

[32] H. Zender, P. Jensfelt, Ó. Martinez Mozos, G.-J. M. Kruijff, and W. Burgard, 'An integrated robotic system for spatial understanding and situated interaction in indoor environments', in *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI 2007)*, (July 2007).

[33] P. Ziaie, T. Müller, M. E. Foster, and A. Knoll, 'Using a naïve Bayes classifier based on k-nearest neighbors with distance weighting for static hand-gesture recognition in a human-robot dialog system', in *Proceedings of CSICC 2008*, (March 2008).