Integrating Language, Vision and Action for Human Robot Dialog Systems

Markus Rickert, Mary Ellen Foster, Manuel Giuliani, Tomas By, Giorgio Panin, and Alois Knoll

Robotics and Embedded Systems Group
Department of Informatics, Technische Universität München
Boltzmannstraße 3, D-85748 Garching bei München, Germany
{rickert,foster,giuliani,by,panin,knoll}@in.tum.de

Abstract. Developing a robot system that can interact directly with a human instructor in a natural way requires not only highly-skilled sensorimotor coordination and action planning on the part of the robot, but also the ability to understand and communicate with a human being in many modalities. A typical application of such a system is interactive assembly for construction tasks. A human communicator sharing a common view of the work area with the robot system instructs the latter by speaking to it in the same way that he would communicate with a human partner.

1 Introduction

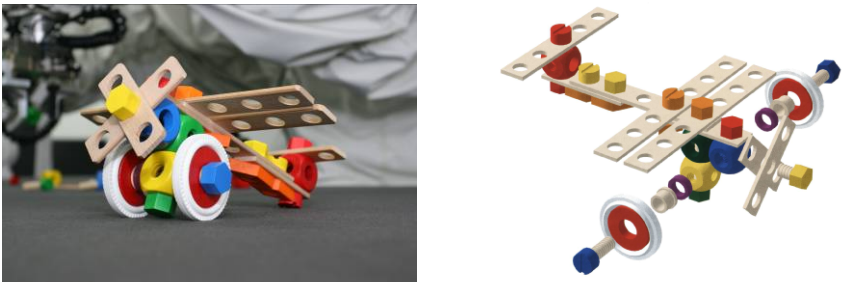
Joint action between humans and robots of various bodily structures (i.e., *N* robots with *M* humans) can be effectively supported by carrying on dialogs in natural language, by interpreting/generating facial expressions and gestures and by direct physical cooperation. It is not difficult to predict that the development of techniques for effective and efficient joint action based on multimodal communication flow between humans and artifacts is one of the most important requirements for the advancement of humanoid and service robotics in the wide sense. Once this development gains enough momentum, the requirements for ever-increasing responsiveness of the robots, for wider applicability to new scenarios, situations and object domains, and for an easy integration of the robots into scenarios with many cooperating robots and many cooperating humans will grow quickly.

There have been various attempts to design robots that directly interact with humans—for the purpose of "programming by demonstration" [1], for controlling their behavior within certain limits, or for "force amplification". These approaches, while working under controlled laboratory conditions, have not met with much success. We identify two main reasons for this failure: (i) Human instructions are perceived by the robot in just one modality, mostly through a fixed camera. This prevents the system from constructing cross-modal associations by evaluating clues from other modalities (sound, touch, etc.). It also prevents humans from giving additional explanations in *natural* modalities, e.g. combining hand movements with speech to give instructions. (ii) Partly due to mono-modality, the communication flow

C. Stephanidis (Ed.): Universal Access in HCI, Part II, HCII 2007, LNCS 4555, pp. 987–995, 2007.

for enabling joint action is not in the form of a dialog between human and robot. Such dialog becomes indispensable in the case of error conditions. Furthermore, the aspect of physical cooperation between the human and the robot system for the purpose of *instructing* the robot has rarely been addressed.

From a cognitive perspective, it is highly desirable that the robot systems autonomously realize cognitively adequate modes of interaction with humans— for dialog control, for synchronizing utterances with motor control and for lifelong learning and plasticity. To show how humans may communicate with robot systems naturally and multimodally, in a large-scale research effort, we are developing a system for joint action between robots and humans integrating vision, language, and



(a) (b)

Fig. 1. The Baufix aircraft toy scenario

Fig. 2. Setup with two industrial robot manipulators and an animatronic head

cooperative action. To the best of our knowledge it is the only system that realizes joint action between humans and robots as well as between robots and that is controlled through multimodal dialog (primarily complex language utterances accompanied by gestures).

The scenario chosen for our system is the joint construction process of wooden toy—model of an aircraft or similar objects (Fig. 1). A human instructor with robot partner cooperates to build aggregates from elements of the toy construction set [2].

The setup (Fig. 2) consists of two industrial robot manipulators mounted to resemble human arms and an animatronic head [3] so that the user can interact with the system across a table. Various cameras are installed for the recognition of objects on the table and for tracking the user. Force/torque sensors and grippers (with position sensors) on each arm complete the installation.

2 Related Work

This system is partly based on previous work carried out at the University of Bielefeld [4,5,6]. These robots were controlled by a negotiating multi-agent system [7] that autonomously combined a large set of parametrized action primitives into action sequences based on the instruction and the environmental conditions. This system performed quite well given the state of the technology ten years ago. However, it did not have overly powerful recognition abilities, computer vision was slow, and so was the overall operation speed.

Several more recent systems have also considered the task of human-robot collaboration. Leonardo [8], for example, is a fully-embodied humanoid robot with social skills that allow it to learn and collaborate effectively in human settings. Mel the robotic penguin [9] 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 [10] 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 in dealing with a situation. The Karlsruhe humanoid robot [11] supports dialog-based human-robot collaboration in a kitchen environment.

3 Architecture

The system's architecture (Fig. 3) is divided into five high-level components, each with several functional modules. The data sent between the different modules is given timestamp to enable multimodal integration. Several computers are used to balance the processing load: for example, separate machines are used for object recognition and robot control. Inter-module communication is implemented using the Internet Communications Engine (Ice) [12], an object-oriented middleware which supports distributed heterogeneous systems. In this section we will give an overview of the different components and explain the connections between them.

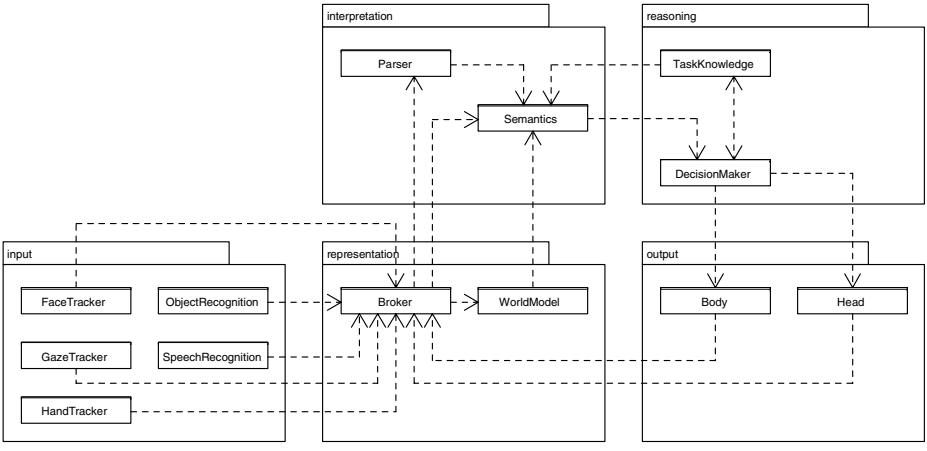


Fig. 3. Overview of the system's architecture

3.1 Input

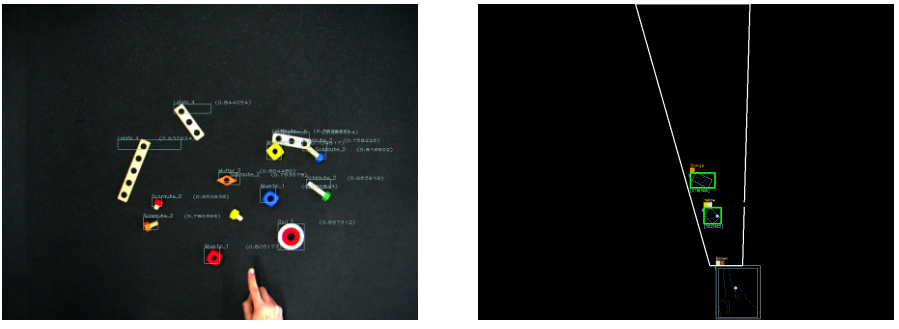
In order to enable human-robot interaction with an emphasis on joint-action we require several input modalities. Apart from a focus on speech and verbal communication, we also support several non-verbal input channels.

The *SpeechRecognition* module is based on the Dragon NaturallySpeaking software [13], and provides speaker-independent limited-vocabulary speech recognition. The speech recognition hypotheses are passed on to the *Parser* module for further processing, as described in Section 3.2.

An implementation of template-matching based object recognition in combination with tracking is the core of the *ObjectRecognition* module, which is responsible for identifying the toy parts on the table. The position and orientation in world coordinates, as well as the type and color of the objects are determined using calibrated camera mounted on top of the table (Fig. 4(a)). The module is also able to detect overlapping objects to some degree. The use of template-matching will later provide the opportunity to introduce new templates into the system at run-time: if the system is unable to recognize an object, it can ask the user to identify it. The learned template is then stored for further use. The current implementation of this module also provides an early realization of the *HandTracking* module's functionality (Fig. 4(b)), which enables the user to refer to objects by pointing at them.

The *FaceTracker* system currently in use estimates the head position in space by tracking the external contour (Fig. 5(a)). It uses the Contracting Curve Density (CCD) algorithm, which has been improved for real time applications in [14]. The animatronic head uses this information in order to address the user during conversation.

The next version currently under development will provide a full 6 DOF estimation of the head position and orientation (Fig. 5(b)), together with an estimation of the user's gaze direction (Fig. 5(b)) for the *GazeTracker* module. For this purpose, we



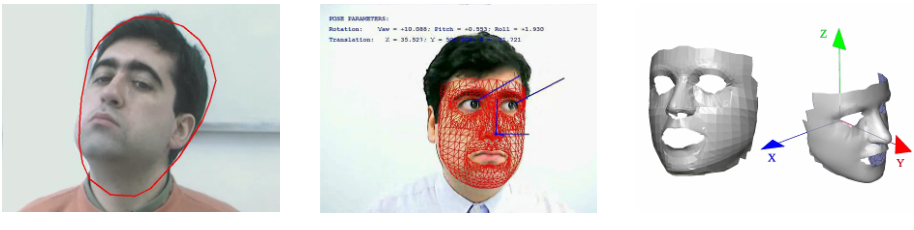
(a) (b)

Fig. 4. Object and hand recognition

adapt generic head model (Fig. 5(c)) [15] by fitting it to specific person's face using two photos, one from the front and one from the side view. After the initial head contour pose estimation, the 3D face template is matched to the underlying image by using fast Mutual Information optimization [16]. This provides more robust approach to deal with variations between expected and observed appearance due to changing lighting conditions, occlusions and other nonlinear effects. Two cameras, one mounted at the robot's base (gaze) and one just below the head (face) provide the input for these modules.

3.2 Representation

This part of the system is responsible for the symbolic representation of the different input channels. A core part in the communication of the system is provided by the *Broker* module. All information processed by the input and output modules is directed to this module, and other modules can register for the event types they are interested in. The data is then passed on to the relevant modules using Ice's publish/subscribe mechanism. This module also provides facilities for centralized logging of system activity for use in debugging and system evaluation.



(a) (b) (c)

Fig. 5. Face tracking with external contour position and 6 DOF estimation

The *WorldModel* module provides a constantly-updated abstract representation of the current set of Baufix components in the world and their physical configuration.

3.3 Interpretation

The *Parser* module uses grammar to process utterances in natural language. It is based on the Combinatory Categorial Grammer (CCG), combination of regular categorial grammars and the combinatory logic [17]. This grammar is able to parse imperative sentences as well as questions, statements and confirmations. It uses the OpenCCG implementation [18] and currently supports sentences in the German language; an English grammar is under development.

The *Semantics* module responds to processed input from the speech recognizer and parser, and uses information from the world model and the other multimodal input channels to select the most likely interpretation of the user's requests. This interpretation is then passed to the *DecisionMaker* component described in the next section. If the user input cannot be fully interpreted—for example, because particular words in the speech were not understood correctly, if the multimodal input is unclear, or if reference cannot be unambiguously resolved—the interpreter does as much processing as it can and passes the result along.

3.4 Reasoning

The *TaskKnowledge* module creates plans for assembling the aircraft, but only abstractly, without considering the physical geometry of the pieces, arms, grippers etc. A "plan" in this sense is data structure in this module, and there can be any number of plans held in memory simultaneously, for example alternative suggestions for building particular aircraft. In addition, the system has notion of the "current plan" (for building the aircraft) which has been agreed on by the system and the user. At any given time, part of this plan has already been acted on, and the rest is to be done in the future. As part of the normal input processing, the system matches the perceived actions and utterances of the user against the current plan, so that the plan may advance without any action on the part of the robot.

The *DecisionMaker* module responds to interpreted messages from the input processing system and selects the appropriate system response. This decision is based on both the user input and the current state of the construction task. In the case of partially resolved input, the dialog manager can respond in one of two ways. It can either ask the user to clarify or repeat the request until it is fully understood, or it can make a guess as to the most likely interpretation and rely on the user for correction if it is wrong. The former strategy is less error-prone, but may lead to user frustration; the latter is based on the observed behavior of humans cooperating with other humans. As we discuss in Section 4, we intend to compare these two strategies in user evaluation.

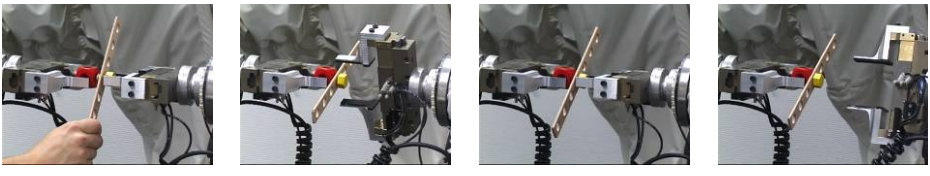
3.5 Output

There are two main output channels in the system: actions of the robot manipulators, and embodied speech of the robot head. All output on both channels is sent to the Broker module for logging and so that it can be used to interpret utterances from the user.

The possible actions of the robot body are abstracted through set of action behaviors [19,20,21,22]. Parameterized sensorimotor actions can be chained together

to customize new behaviors. These are executed on the robot using online real-time control instead of generating offline trajectories. Because of the live interaction with the human instructor, the system needs to be able to abort any action instantly and safely to provide "barge-in".

With given attractor position/force, the corresponding trajectory for the controller needs to be calculated online and the robot's velocity/acceleration constraints have to be considered. Different exit conditions, optionally combined with boolean operators, can be used to determine the next operation in this finite state machine. The screw action requires such complex combination to implement spiral search pattern for compensation, re-grasping due to cable limitations and the determination of the final force/torque limits. Fig. 6 shows such joint-action operation where the user is helping the robot with the assembly by holding slat.



(a) (b) (c) (d)

Fig. 6. Human and robot cooperating in an assembly

With online trajectory generation, a signal from the *DecisionMaker* module can activate the calculation of stop trajectory or the seamless switch to the next action at any time. The information from the robot's joint positions as well as from the gripper's position sensor are sent to the Broker module for further use. The *ObjectRecognition* can for example utilize this data to remove the manipulators from the top view image. Based on feedback from the gripper, the *DecisionMaker* can also determine errors during various actions in addition to information from the object recognition. If the goal was to pick up cube and the closed span does not match or indicates full close, then wrong object or no object has been picked up.

The linguistic content of the synthesized speech is created by OpenCCG, using the same OpenCCG-based grammar as the Parser module. The speech is synthesized using AT&T NaturalVoices [23] and is accompanied by lip-synchronization and other non-verbal behaviors of the animatronic head. Other non-verbal behaviors include gazing at the user while talking to them (using the position information from the face tracker), looking at objects on the table while manipulating them, and using nodding and other facial expressions to support the content of the speech.

4 Future Work

Future work for the JAST human-robot dialog system falls into two main categories: evaluating the current version of the system and continuing development of the next version. For the evaluation, we intend to investigate two aspects of the system: the non-verbal behavior of the talking head, and the dialog strategies used to respond to imperfectly understood output.

Previous studies with embodied agents [24] have shown that gaze feedback can increase engagement between human and an embodied robot agent; however, other studies with embodied agents have indicated that an expressive talking head may actually harm task performance in some cases [25]. We will employ widely-used paradigm to evaluate the impact of the non-verbal agent behavior on the interaction quality. Users will interact with the system in one of two modes: one in which the head uses its full repertoire of non-verbal gaze and expressive behavior, and one in which all behaviors except for lip-synchronization are disabled. The evaluation will also compare the impact on the dialog of the two strategies for responding to ambiguous output: waiting for complete clarification, or choosing an option arbitrarily and relying on the user for correction if necessary.

For the next version of the system, we aim to expand its technical capabilities in all areas. For input processing, we will continue development on the objectrecognition, gaze-tracking and hand-tracking modules; for the interpretation and reasoning modules, we aim to integrate more sophisticated assembly-planning knowledge and more full-featured dialog manager based on the Information State Update approach [26]; while on the output side, we will extend the robot's repertoire of actions and enhance the non-verbal displays of the talking head.

We will also expand the linguistic processing modules so that the system is able to work in English in addition to German.

Acknowledgments. This work was supported by the EU FP6 IST Cognitive Systems Integrated Project "JAST" (FP6-003747-IP), http://www.euprojects-jast.net/.

References

1. Erlhagen, W., Mukovskiy, A., Bicho, E., Panin, G., Kiss, C., Knoll, A., van Schie, H., Bekkering, H.: Goal-directed imitation for robots: a bio-inspired approach to action understanding and skill learning. Robotics and Autonomous Systems 54(5), 353–360 (2006)
2. Foster, M.E., By, T., Rickert, M., Knoll, A.: Human-robot dialogue for joint construction tasks. In: Proceedings of the International Conference on Multimodal Interfaces, pp. 68–71. ACM Press, New York (2006)
3. van Breemen, A.J.N., Yan, X., Meerbeek, B.: iCat: An animated user-interface robot with personality. In: Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 143–144. ACM Press, New York (2005)
4. Knoll, A., Hildebrandt, B., Zhang, J.: Instructing cooperating assembly robots through situated dialogues in natural language. In: Proceedings of the IEEE International Conference on Robotics and Automation (April 1997)
5. Zhang, J., Knoll, A.: A two-arm situated artificial communicator for human-robot cooperative assembly. In: Proceedings of the IEEE International Workshop on Human Robot Communication, pp. 292–299 (2001)
6. Knoll, A.: basic system for multimodal robot instruction. In: Kühnlein, P., Rieser, H., Zeevat, H. (eds.) Perspectives on Dialogue in the New Millennium. Pragmatics and Beyond New Series, vol. 114, John Benjamins, Amsterdam (2003)
7. Knoll, A.: Distributed contract networks of sensor agents with adaptive reconfiguration: Modelling, simulation, implementation. Journal of the Franklin Institute 338(6), 669–705 (September 2001)

8. Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., Lieberman, J., Lockerd, A., Chilongo, D.: Tutelage and collaboration for humanoid robots. International Journal of Humanoid Robotics 1(2), 315–348 (2004)
9. Sidner, C.L., Dzikovska, M.: first experiment in engagement for human-robot interaction in hosting activities. In: Bernsen, N., Dybkjær, L., van Kuppevelt, J. (eds.) Advances in Natural Multimodal Dialogue Systems, Springer, Heidelberg (2005)
10. Fong, T.W., Kunz, C., Hiatt, L., Bugajska, M.: The human-robot interaction operating system. In: Proceedings of the International Conference on Human-Robot Interaction, ACM (2006)
11. Burghart, C., Mikut, R., Stiefelhagen, R., Asfour, T., Holzapfel, H., Steinhaus, P., Dillmann, R.: cognitive architecture for humanoid robot: first approach. In: Proceedings of the IEEE-RAS International Conference on Humanoid Robots, pp. 357–362 (2005)
12. Henning, M.: new approach to object-oriented middleware. IEEE Internet Computing 8(1), 66–75 (2004)
13. Nuance Communications, Inc.: Dragon NaturallySpeaking http://www.nuance. com/ naturallyspeaking/
14. Panin, G., Ladikos, A., Knoll, A.: An efficient and robust real-time contour tracking system. In: Proceedings of IEEE International Conference on Computer Vision Systems, IEEE Computer Society, 44 (2006)
15. Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., Salesin, D.H.: Synthesizing realistic facial expressions from photographs. In: Proceedings of the International Conference on Computer Graphics and Interactive Techniques, pp. 75–84. ACM Press, New York (1998)
16. Moddemeijer, R.: On estimation of entropy and mutual information of continuous distributions. Signal Processing 16(3), 233–246 (1989)
17. Steedman, M.: The Syntactic Process. MIT Press, Cambridge, MA (2000)
18. White, M.: Efficient realization of coordinate structures in combinatory categorial grammar. Research on Language and Computation 4(1), 39–75 (2006)
19. Morrow, J.D., Khosla, P.K.: Manipulation task primitives for composing robot skills. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 3354–3359 (1997)
20. Zhang, J., von Collani, Y., Knoll, A.: Interactive assembly by a two-arm robot agent. Journal of Robotics and Autonomous Systems 29(1), 91–100 (1999)
21. Thomas, U., Finkemeyer, B., Kröger, T., Wahl, F.: Error-tolerant execution of complex robot tasks based on skill primitives. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 3069–3075 (2003)
22. Milighetti, G., Kuntze, H.-B.: Multi-sensor controlled skills for humanoid robots. In: Proceedings of the IFAC International Symposium on Robot Control (2006)
23. Jilka, M., Syrdal, A.K.: The AT&T German text-to-speech system: Realistic linguistic description. In: Proceedings of the International Conference on Spoken Language Processing, pp. 113–116 (2002)
24. Sidner, C.L., Lee, C., Kidd, C.D., Lesh, N., Rich, C.: Explorations in engagement for humans and robots. Artificial Intelligence 166(1–2), 140–164 (2005)
25. White, M., Foster, M.E., Oberlander, J., Brown, A.: Using facial feedback to enhance turn-taking in multimodal dialogue system. In: Proceedings of the International Conference on Human-Computer Interaction (2005)
26. Traum, D., Larsson, S.: The information state approach to dialogue management. In: Smith, R.W., van Kuppevelt, J. (eds.) Current and New Directions in Discourse and Dialogue, Kluwer Academic Publishers, Dordrecht (2003)