

Realization of Natural Interaction Dialogs in Public Environments using the Humanoid Robot ROMAN

Norbert Schmitz, Jochen Hirth, and Karsten Berns

*Robotics Research Lab, Department of Computer Science
University of Kaiserslautern, Gottlieb-Daimler-Strasse
67663 Kaiserslautern, Germany
{nschmitz, j_hirth, berns}@informatik.uni-kl.de*

Abstract— The ability of humanoid robots to perform naturally with multi-modal dialogs in an unmodified environment is very limited. Although various solutions for sub-problems of interaction like tracking or speech recognition exist, it is obviously a great challenge to integrate all the requirements into one common platform. This paper presents experiments with the humanoid robot, ROMAN, performed in an unchanged environment with public access. A short dialog with interaction possibilities is chosen and the experimental results are described.

I. INTRODUCTION

Research activities in the field of humanoid robotics often claim to realize a natural interacting robot. In contrast to robots the interaction possibilities of humans are very complex and a complete implementation is currently not possible. To overcome this problem it is important to limit the sensors, actors and the control architecture to a less complex and realizable scenario. Another way to reduce the complexity is the usage of a virtual agent and the modification of the environment. Since neither virtual agents nor modified environments are desirable for a humanoid robot, some experiments with the embodied agent ROMAN, within an unmodified public environment and a predefined dialog situation, are performed.

Virtual agents like the humanoid MAX are intended to realize multimodal communication with a humanoid interaction partner. In [1] a turn taking scenario is described. The robot uses virtual sensors which perfectly provide all necessary information about the environment. In contrast to this example this paper presents an embodied agent – the humanoid robot ROMAN – with a real sensor system. With this equipment it is possible but much more complex to realize a communication scenario without any modifications to the environment.

In contrast to virtual agents experiments with embodied robots are often less complex. In [2] the humanoid robot Maggie is described in a dancing scenario. In this scenario the robot reacts on the contact of robot and human using tactile sensors. The described setting requires the user to know how the communication can be realized which limits the communication to the predefined set of signals.

The humanoid robot QRIO is able to perform more complex dialogs using the perceptions of cameras and microphones [3] embedded in the EGO control architecture. Although the interaction process shows the possibilities of current communication systems, a more natural communication in public

environments can only be realized with a normal sized robot.

In contrast to full body agents, the robot MERTZ [4] consists only of a humanoid head. Experiments in office and public environments show that the interaction without a predefined dialog, using special toys and a mimicking behavior, is possible. Although the interaction with an artificial robot is promising, a human-like agent could present much more natural and well known interaction signals.

The humanoid robot BARTHOC is equipped similar to ROMAN and also able to perform multimodal dialogs. In [5] experimental results with the robot are evaluated concerning the tracking and memory of possible interaction partners. Although promising results are shown, the robot has not been tested in public environments.

II. THE HUMANOID ROBOT ROMAN

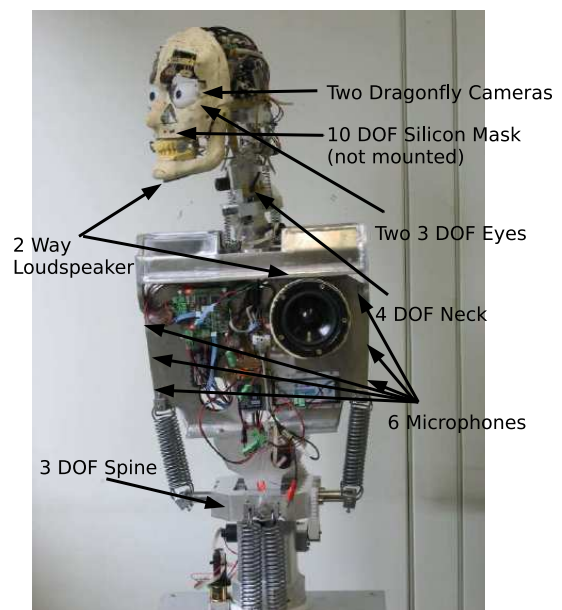


Fig. 1. Image of the robot ROMAN with description of sensors and actors.

The humanoid robot ROMAN (see figure II) currently has 24DOF (degrees of freedom): 3DOF in the lower part of the upper body, a neck with 3DOF similar to the one in the body, an additional neck joint in the head for rotations over the

horizontal axis, the eyes with two rotations over horizontal and vertical axis as well as the movements of the eyelid, 1DOF for the movements of the lower jaw and 10DOF for the movements of the silicon mask.

The sensor system consists of 2 Point Grey Dragonfly cameras located in the eyeballs with a resolution of 640x480 pixels and a frame rate of up to 30 fps, 6 microphones are located in the upper body and 4 infrared distance sensors are mounted on the body.

The motor control is realized using 4 Freescale DSPs and the control architecture is running on two standard personal computers.

III. EMOTION-BASED ARCHITECTURE

The goal of the ROMAN project is to realize natural human-robot interaction, therefore the robot ROMAN needs the ability to perform non-verbal as well as verbal interaction. Since psychologists point out that approx. 60%–70% of the human-human interaction is conducted non-verbally [6], therefore it is of enormous importance to enable robots to use these information.

To realize this goal, ROMAN requires the abilities to express non-verbal signals in a way that they can be recognized and correctly interpreted by humans and, on the other hand, to recognize non-verbal signals expressed by humans. Most of the non-verbal signals of humans depend on their emotional state and are activated in their subconscious mind. These expressions are not limited to gestures or mimic, the whole body generates non-verbal signals [7]. Based on these facts a robot which is able to perform non-verbal human-robot interaction needs some kind of emotional state.

The emotion-based control architecture of the robot is designed in accordance to the psychological theories in [8]. The whole architecture is shown in figure 2. The control architecture consists of 4 main groups: motives, emotional state, habits of interaction, and percepts of interaction.

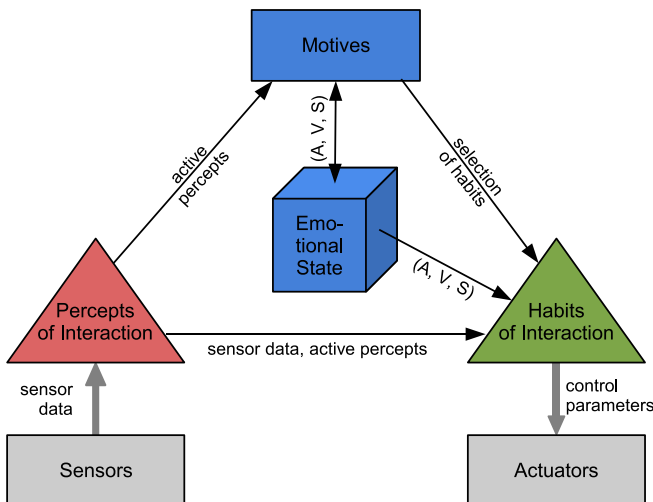


Fig. 2. The emotion-based control architecture of the humanoid robot head ROMAN. The emotional state of the system is represented by the 3 dimensions arousal (A), valence (V), and stance (S).

For the representation of the emotional state, a 3-dimensional space according to [9] with arousal, valence, and stance (A, V, S) as axes is selected. The emotional state influences the motives in their processing of the percept information as well as the habits of interaction. In different emotional states, the robot will interpret the information of the environment in different ways and will also act in different ways. For example if it is very nervous, the movements will be executed much faster and also its speech will be conducted much faster in comparison to the neutral state. Besides this, the actual emotion is displayed by facial expressions and body postures.

For the realization of habits a system for the description and implementation of interaction signals —Habits of Interaction (HI)— was developed [10]. As basic modules for the HI the behavior modules of iB2C¹ (integrated Behavior-Based Control), figure 3, are used. This module has 3 inputs: stimulation s , inhibition i and data input \vec{e} , and 3 outputs: activity a , target rating r and data output \vec{u} . In addition, this module has a function for the calculation of the output data depending on the input: $F(\vec{e}, \iota, i) = \vec{u}$. This function is called transfer function. The activation ι of a behavior is calculated depending on the stimulation and the inhibition: $\iota = s \cdot (1 - i)$. For the realization of complex HI, e.g. emotional expressions, a hierarchy of HI is developed which enables the robot to generate complex HI as a combination of basic ones. HI concerning the same or similar body parts can be grouped. Based on the hierarchical ordering of the HI, different complexity levels can be generated. That way a complex behavior network for the generation of social behavior in an interaction situation is generated.

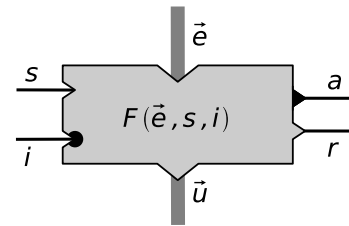


Fig. 3. The iB2C module: all habits of interaction are build out of these modules.

Having the abilities to express non-verbal signals, the robot needs a motivating component that causes the robot to activate behavior sequences. Therefore the motives are realized in the emotion-based architecture. A.E. Kelley [11] points out that the emotional part is crucial for the motivation of human behavior. Because of this the robot needs a component that combines emotional as well as cognitive characteristics. Looking to psychology, for solving this problem, leads to the human motives. As pointed out in [7] motives generate goal directed behavior sequences. A certain motive is defined by its activation, direction, intensity, and duration. They represent the combination of emotions, needs, and their satisfaction.

For the implementation of the motives, again, the behavior

¹<http://rrlib.cs.uni-kl.de/>

TABLE I
THE REALIZED MOTIVES OF THE ROBOT ROMAN AND THEIR FUNCTION

Motive	Function
Obeys humans	If a human gives the robot an order to do something it will stop its actual work and will obey the order.
Self-protection	Generates an evasive movement, if a close object is detected.
Energy consumption	If the robot's energy is low it will try to get new energy, by telling humans in its surrounding.
Avoid fatigue	If the load is too high this motive avoids the starting of new processes.
Communication	If a person is detected, this motive tries to start a conversation and takes care that the person is focused.
Exploration	If the robot is getting bored because of the absence of stimuli, this motive starts the exploration of the robots surrounding.
Entertainment	If ROMAN is exploring its environment but does not get any interesting stimulus, this motive gets active. ROMAN starts to sing a song especially to attract the interest of humans in its surrounding.

based concept of iB2C is used. The motives in the emotion-based architecture of ROMAN, calculate their satisfaction depending on the active percepts of interaction. Depending on this satisfaction every motive influences the actual emotional state. The motives also try to influence the robot's behavior in a way that a satisfied state is reached again. The output of different motives—the change of the emotional state and the activation of habits of interaction—is merged depending on the satisfaction of the motives. The lower the satisfaction of a motive, the higher is its influence on the fusion. In a robot system motives can be used to define some basic goals of the robot [12]. For the testing scenario explained in the remainder of this paper, several motives are implemented. The motives and their function are displayed in table I.

For the verbal human-robot communication as well as for the deliberately generated non-verbal expressions the dialog system explained in [13] is used. Therefore the dialog description is handed over to the speech engine, figure 4.

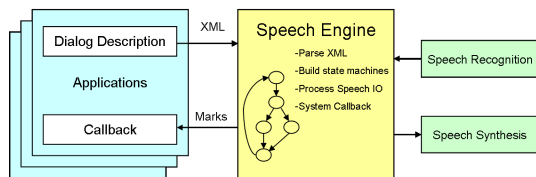


Fig. 4. Architecture of Dynamic Speech System

Dialogs are specified in XML. The XML-scheme includes elements for describing speech inputs and outputs, specifying speech parameters, and defining condition and value variables for controlling and adapting the dialog flow. Additionally it also allows the definition of non-verbal signals which can be recognized or generated at a certain state in the dialog flow. During dialog registration, the given XML file is parsed and mapped to a state machine. In the main dialog section, user inputs which have to be recognized by the speech recognition system or by the camera system and machine outputs which have to be performed by the speech synthesis or by the robots movement system are specified. For example if the system waits for the input "yes", the human dialog partner can either say "yes" or nod. In order to change ROMAN's emotional state or to activate behaviors depending on the actual spoken or recognized words, information can be send back to the robot using the "callback" methods of the speech engine.

For speech synthesis *Profivox* [14], is used. The speech recognition is realized by *AmiASR*. This software is based on the *Hidden Markov Model Toolkit (HTK)*² [15].

The perception system for the humanoid robot fulfills the task of sensor data abstraction. Sensors like cameras and microphones are constantly providing a large amount of data. This data must be used to extract all required information about the environment of the robot which are necessary for an interaction scenario. The selection of relevant and necessary information is complex since it is not possible to represent all aspects of human-human communication. For the humanoid robot ROMAN the perception system is divided into three basic groups: information-capturing, -memory and -extraction. These groups can be developed and integrated independently and allow a modular system design. The capturing recognizes simple information captured from the raw sensor data and combines them to the model of the environment. The memory module stores these models over time to allow a comparison with previous capturing. Finally the extraction groups uses this "time-memory" to extract time dependent information and transfers the information to the motives and habits. Figure 5 shows the three basic parts and the information flow between the modules.

The information capturing system uses the data from the physical sensors, divides them into several "virtual sensors" and combines the knowledge into a simple model of the environment. These models are then passed to the information memory. The notion of "virtual sensor" has been introduced to describe a module which analyzes the real sensor information in respect to a specific property. This concept allows a modular integration of additional sensors for a specific type of information. For the dialog presented in the experiments the following virtual sensors are required:

Face Detector: The face detector uses a trained haar cascade classifier to localize frontal faces in an image. It provides information about size and position of a face.

²<http://htk.eng.cam.ac.uk>

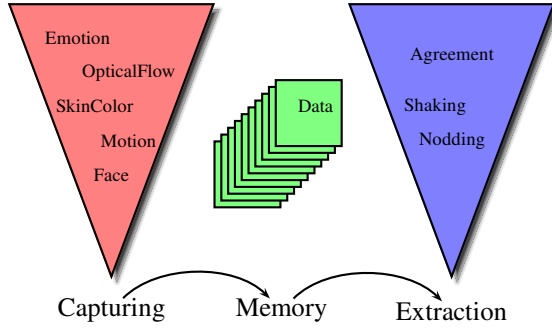


Fig. 5. The perception system of the robot with the basic capturing group, the short term memory module and the high-level information extraction group. The data flow is always directed from capturing to extraction.

Skin Color Detector: The skin color detector assigns a probability value of skin color to each pixel. The classification is based on manually classified training images and a fitting of Gaussians into the RGB color space. The probabilities for each pixel can be obtained using a look up table. [16]

Optical Flow Detector: This virtual sensor calculates the optical flow between two successive images. It provides information about the direction and distance of movement in an image.

The result of the information capturing is a list of candidates for face with related average skin color and average direction of motion. Besides the described information several other information like emotions and gestures are detected but not used for the interaction situation presented in this paper.

The basic information of the capturing system cannot be used directly to extract high-level information. False detections and the missing connectivity between sensor information and object of interest requires the introduction of a multi-object filter. This filtering is realized using a bootstrap particle filter on the image plane. The output of the filter and the selected information from the virtual sensors are combined to a general description of every object in focus of the robot. Currently these objects are limited to faces since they are of main importance for any interaction scenario.

The information memory consist of a short term memory storing the 100 most recent capturings. One element of the memory itself is a vector with up to 10 elements. This enables simultaneous tracking of multiple objects. Objects as described here are currently faces but can be any other object of interest like hands or toys. The memory is internally designed as a ring buffer to reduce the time for memory allocation. Each element stores information of type, state and the corresponding time stamp.

The information extraction group uses the data stored in the short term memory to extract time dependent high level information. This extraction is necessary since actions like nodding, head shaking or hand waving are hard or impossible

to detect in a single image. Another challenge for the information extraction is the question which behaviors are important and possible to detect.

Each of the extracted behaviors is realized as a behavior-based module according to the iB^2C -framework. The activities of the extraction modules represent the probability that the described behavior has been detected in the memory. Two examples of extraction modules are nodding and head shaking which can be used to realize a non-verbal dialog using the vision based detection of agreement or disagreement respectively.

The activity a of the module nodding for example is calculated by

$$a = s \cdot \| (ampl_y - ampl_x) \cdot 10 \| \cdot \| ampl_y \| \quad (1)$$

while $\| \bullet \|$ indicates a limitation to the interval $[0, 1]$ and $ampl_x, ampl_y$ are the average amplitude of the optical flow in x and y direction. The activity of the module nodding is high when the amplitude in y -direction is high and the difference between the amplitude in y - and x -direction is positive and high. These heuristics describe the typical observations when a person is nodding. The head shaking behavior is implemented similar to the nodding behavior with inverted amplitudes in x - and y -direction.

At the end of the information extraction the perception module provides information about position, direction of motion, probability of a human and the activity of nodding and shaking related to the interaction partner. Currently it is assumed that only one interaction partner is present.

IV. EXPERIMENT

To test and verify the implemented architecture experiments were performed in which ROMAN interacts with a human partner³. The experimental setup as well as the robot behavior achieved with the proposed architecture is described in the following section.

The robot ROMAN introduces the Robotics Research Lab to the people walking by. Therefore ROMAN searches for humans, welcome them and ask whether they are interested in information or not. If the human is interested in information ROMAN explains a poster of the work group that is mounted on the wall at the left hand side of the robot. To achieve this goal the following components of the emotion-based architecture are mainly involved. The involved percepts of interaction are the percept for the detection of humans, the percept for the detection of non-verbal signals nodding and shaking the head and the percepts for detection of the distance of a human. The mainly used habits of interaction are the habits for focusing and following humans as well as the dialog system which is involved to generate the verbal communication. To detect humans in the surrounding of the robot the exploration motive is needed. When a person is detected the interaction is triggered by the communication

³A video of the experiment can be found on: <http://agrosy.informatik.uni-kl.de/en/robot-gallery/roman/>

motive. A flowchart of one part of the interaction scenario is displayed in figure 6.

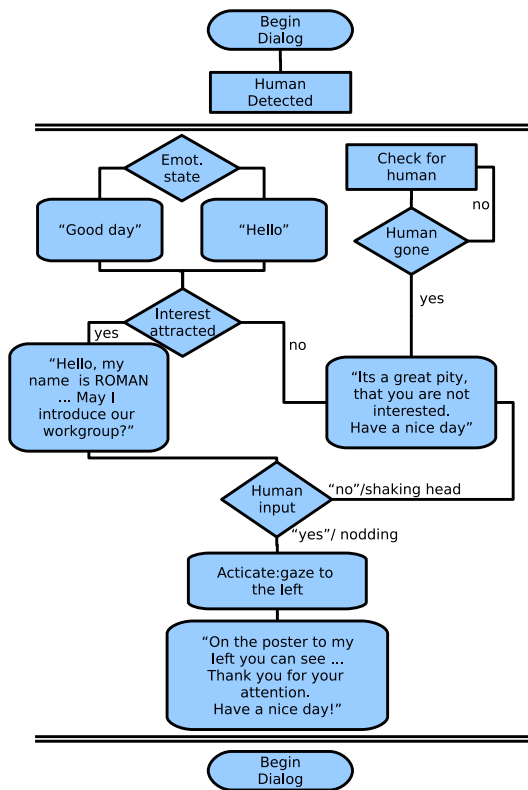


Fig. 6. A flowchart of ROMAN's behavior achieved with the emotion-based architecture during the proposed interaction scenario.

The achieved behavior of ROMAN is as follows: The robot is looking around and searching for humans. Once a human is detected ROMAN tries to attract the human's attention by talking to the human, figure 7. The human shows interest by answering. If the human doesn't show interest the ROMAN says: "It's a pity" and searches for other humans. If the human shows interest, ROMAN introduces itself and asks whether the human is interested on some more information on the Robotics Research Lab. To this question the human can answer by nodding or by saying "yes", or by shaking the head or saying "no".

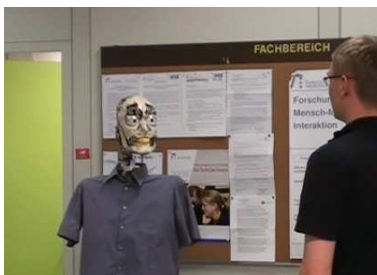


Fig. 7. A typical situation in the experiment, ROMAN asks a pedestrian if he is interested in information on the Robotics Research Lab.

If the human is interested in more information the robot explains a poster of lab, figure 8. Afterwards ROMAN thanks

for the attention of the human and says good bye. If the human is not interested in additional information ROMAN says good bye and looks for other humans. If the human moves away during the interaction the robot says: "It's a pity" and looks for new stimuli. If the human steps to close towards ROMAN during the interaction, ROMAN will stop its actual work and ask the human to step back. Afterwards ROMAN continues the interrupted dialog.

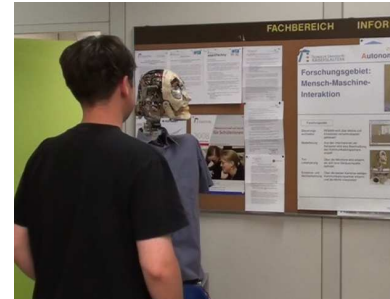


Fig. 8. ROMAN explains a pedestrian the poster of the Robotics Research Lab on its left.

Two typical dialog situations that appeared during this experiment are listed below:

Situation 1:

ROMAN detects a human
 ROMAN: "Hello"
 Human: (steps closer)
 ROMAN: "May I introduce my..."
 Human: (human is too close)
 ROMAN: "Would you please step back!"
 Human: (human steps back)
 ROMAN: "Thank you. May I introduce myself? I'm ROMAN the humanoid robot of the University of Kaiserslautern. I'm able to use non-verbal expressions for interaction. Are you interested in some more information on our work group?"
 Human: "Yes!", (nodding)
 ROMAN: "On the poster to my left (gaze towards the poster) you find more information on our work group. Our work group Thank you for your attention. Have a nice day. Good bye!"
 Human: "Thank you, good bye"

Situation 2:

ROMAN detects a human
 ROMAN: "How do you do?"
 Human: (steps closer)
 ROMAN: "May I introduce myself. I'm ROMAN the humanoid robot of the University of Kaiserslautern. I'm able to use non-verbal expressions for interaction. Are you interested in some more information on our work group?"
 Human: "No!", (walks away)
 ROMAN: "It's a pity that you are not interested in some information on our work group."

The experimental results from the interaction situation in public environment are quite promising. The robot can detect and focus humans and also interact with them in a simple way. Anyhow there are still some points that are needed to be improved. In order to get the interaction more natural, also the non-functional aspects of human motions also need to be transferred to ROMAN, otherwise the robot appears very inanimate and stiff. Also the detection of humans needs to be improved and expanded. ROMAN sometimes loses a human although he or she is still in the view of the robot. Other sensor systems like microphones should also be used to detect whether somebody is calling for ROMAN or talking to it. After solving these problems, the dialog scenario can be expanded in order to realize it in a more natural way.

During the experiments the behavior of the human interaction partner often is not as “natural” as expected. Humans intend to test what the possibilities of the robot are. The try to step very close just to generate a reaction of robot. This behavior would never be observed within a human-human interaction situation. There are many more examples like extremely aggressive hand waving or testing the face tracking capabilities of the robot. Further experiments should at least to some extent detect these “strange” behavior and react either annoyed or aggressive.

V. CONCLUSIONS

In this paper a natural interaction situation between the robot ROMAN and humans, in public environment, was presented. At first, an overview of the state of the art in natural human-robot interaction in public environment was given and the mechatronic system of ROMAN was described briefly. In the main part of this paper, the components of the emotion-based control architecture involved in the experiment are introduced. The conducted experiment as well as the results are explained in detail. Although there are still some improvements to be made in order to reach a natural human-robot interaction, the results of the experiment are satisfactory. ROMAN is able to track and focus its human interaction partner and ROMAN communicate with the humans using verbal and non-verbal interaction signals. Because of the enormous importance of gestures for the non-verbal interaction the inclusion of arms is planned for the future. This includes the requirement of additional habits of interaction using the arms. Another necessary extension will be the usage of the microphones to detect the position of a speaker, so that ROMAN can also track the interaction partner when the speaker is out of the robot’s view.

Psychological experiments concerning the whole system behavior in human-robot interaction situations must be realized. Also the acceptance of such a robot system by humans needs to be evaluated. Depending on these results the system behavior and appearance can be improved. As final test scenario a situation can be imagined where the robot assists a human solving some kind of puzzle. The robot knows the correct solution of the puzzle and should motivate the human by showing emotional expressions. It should also tell the human the next correct move, if the human asks. The results of

this final test will be interpreted in cooperation with the psychologist involved in the ROMAN project.

ACKNOWLEDGMENT

This work is realized supported by the International Research Training Group of the University of Kaiserslautern (IRTG 1131) which is funded by the Deutsche Forschungsgemeinschaft (DFG).

REFERENCES

- [1] N. Leßmann, A. Kranstedt, and I. Wachsmuth, “Towards a cognitively motivated processing of turn-taking signals for the embodied conversational agent max,” in *Proceedings of the Workshop on Embodied Conversational Agents: Balanced Perception and Action, Conducted at the International Conference on Autonomous Agents and Multiagent Systems (AAMS)*, New York, USA, 2004, pp. 57–64.
- [2] M. Salichs, R. Barber, A. Khamis, M. Malfaz, J. Gorostiza, R. Pacheco, R. Rivas, A. Corrales, E. Delgado, and D. Garcia, “Maggie: A robotic platform for human-robot social interaction,” in *Proceedings of the IEEE International Conference on Robotics, Automation and Mechatronics (RAM)*, Bangkok, Thailand, June 2006.
- [3] K. Aoyama and H. Shimomura, “Real world speech interaction with a humanoid robot on a layered robot behavior control architecture,” in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA)*, Barcelona, Spain, April 18–22 2005, pp. 3825–3830.
- [4] L. Aryananda and J. Weber, “Mertz: A quest for a robust and scalable active vision humanoid head robot,” in *IEEE-RAS/RSJ International Conference on Humanoid Robots (Humanoids 2004)*, 2004.
- [5] T. Spexard, M. Hanheide, and G. Sagerer, “Human-oriented interaction with an anthropomorphic robot,” in *IEEE Transactions on Robotics, Special Issue on Human-Robot Interaction*, December 2007.
- [6] R. Birdwhistell, *Kinesics and Context: Essays in Body Motion Communication*. Philadelphia, USA: University of Pennsylvania Press, 1970.
- [7] H. Hobmair, S. Altenhan, S. Betscher-Ott, W. Dirrigl, W. Gotthardt, and W. Ott, *Psychologie*, H. Hobmair, Ed. Bildungsvverlag EINS, 2003.
- [8] E. Rolls, *The Brain and Emotion*. Oxford University Press, 1999.
- [9] C. Breazeal and R. Brooks, “Robot emotion: A functional perspective,” in *Who Needs Emotions? The Brain Meets the Robot*, J. Fellous and M. Arbib, Eds. Oxford University Press, 2005.
- [10] J. Hirth and K. Berns, “Concept for behavior generation for the humanoid robot head roman based on habits of interaction,” in *Proceedings of the IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, Pittsburgh, USA, November 29–December 1 2007.
- [11] A. Kelley, “Neurochemical networks encoding emotion and motivation – an evolutionary perspective,” in *Who Needs Emotions? – The Brain Meets the Robot*, J. Fellous and M. Arbib, Eds. Oxford University Press, 2005, pp. 29–77.
- [12] J. Hirth and K. Berns, “Motives as intrinsic activation for human-robot interaction,” in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nice, France, September 22–26 2008, pp. 773–778.
- [13] J. Koch, J. Wettach, H. Jung, and K. Berns, “Dynamic speech interaction for robotic agents,” in *13th International Conference on Advanced Robotics (ICAR07)*, Jeju, Korea, August 21–24 2007, pp. 665–670.
- [14] G. Nemeth, G. Olaszy, P. Olaszi, G. Kiss, C. Zainko, and G. Gordos, “Profivox - a hungarian text-to-speech system for telecommunications applications,” *International Journal of Speech Technology*, vol. 3, pp. 201–215, 2000.
- [15] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, J. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Microsoft Corporated, 2000.
- [16] M. Jones and J. Rehg, “Statistical color models with application to skin detection,” *International Journal of Computer Vision*, vol. 46, no. (1), pp. 81–96, 2002.