Perception System for Naturally Interacting Humanoid Robots

Karsten Berns, Norbert Schmitz

The development of humanoid robots rises many challenges in the areas of mechatronics, control and sensor processing. Several partial solutions in the mentioned areas are available but still many unsolved problems do exist. This paper presents a basic concept for a sensor processing system for humanoid robots. The proposed perception system consists of three basic components: information capturing, information memory and information extraction. The capturing module interprets the measurements of the sensors and generates basic statements about the environment. This information is combined to a compact description of objects of interest which is stored in a short term memory. The short term memory is used to extract time depended information from the object descriptions. Experimental results performed with the humanoid robot ROMAN and a behavior-based control architecture show the capabilities of the proposed approach.

1 Introduction

Humanoid robots as accepted members of society with human-like skills can be found in fiction since many years. The demand for such multi-functional robots is high since the application areas are spread over a wide variety of activities like entertainment, house keeping or elderly care. Nevertheless current developments of humanoid robots can only cover small aspects in the development process of such a robot. The mechanical construction of a human-like robot alone is a challenge which could not be completely solved yet and the complex environment of humans poses a great challenge for the control and sensor system. But even with a perfect mechatronic and sensor solution the interpretation and reaction on given environment situations is a highly complex task. Psychologists [9] are analyzing the behavior of interacting humans but the complexity of the communication process does not allow a complete model yet. Nevertheless some reduced models do exist which can be used to improve the behavior of a humanoid robot in a robot-human communication scenario.

The development of the humanoid robot ROMAN (RObot-huMAN interaction machine) at the University of Kaiserslautern focuses on the aspect of natural interaction between humans and robots. The robot as depicted in Fig. 1 shows the current skeleton which can be covered with a silicone skin and clothes. With a height of about 180cm, the skin and the clothes the robot creates the impression of an upright standing human being. Overall the robot has 24 degrees of freedom (DOF): 3 in the spine, 4 in the neck, 3 in each eye and 11 for facial expressions. Further details concerning the mechanical construction can be found in [8].

The control system is implemented using a behavior-based emotion architecture. The emotional space as central object influences the three main parts of the control architecture: habits, motives and percepts. The percepts of interaction will be discussed in detail now whereas further information concerning the habits, motives and the emotional space can be found in [6].

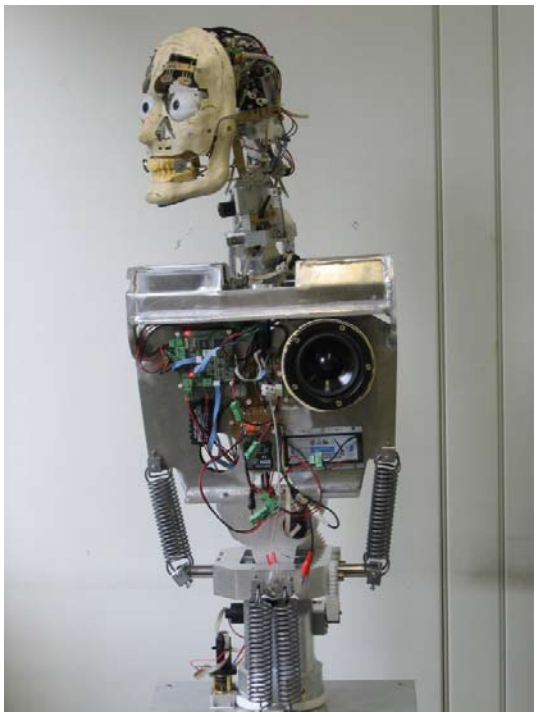

Figure 1: The skeleton of the humanoid robot ROMAN without clothes and silicon mask

2 State of the Art

The manipulating robot ARMAR III [1] is implemented using a hierarchical planning-coordination-execution architecture. The sensor system is not explicitly represented in the architecture so that the information gathered from the sensors is directly used in each layer. In contrast to this approach ROMAN realizes a specific perception module which uses the basic sensor information to generate abstract high-level information to reduce the sensor complexity for the habits and motives.

For any human-robot interaction scenarios a stable approach for tracking humans over time is necessary. Marker

Auszug aus: Künstliche Intelligenz, Heft 1/2009, ISSN 0933-1875, BöttcherIT Verlag, Bremen, **www.kuenstliche-intelligenz.de/order**

less motion capture systems as presented in [2] are well suited for humanoid robots. The use of multiple cues like edges and regions is a promising approach and increases the robustness. This approach does not provide any information how the tracking data is used to control the interaction process. The perception system of the robot ROMAN realizes a similar tracking system with further information about the interaction partner and a mechanism for the extraction of time-dependent information.

The humanoid robot QRIO [5] implements a sensor system with long and short term memory. It basically uses color information like skin and object color gathered from the cameras to influence the interaction process. The concept of capturing and memorizing can also be found in the architecture of ROMAN although the concept of QRIO is extended by the information extraction group.

The benefits of the proposed perception system will be discussed in detail in the following section.

3 Perception System

A perception system for humanoid robot fulfills the challenging 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 inter-human communication.

The interaction process itself should be "natural" which does not allow environmental changes like markers and includes verbal and nonverbal interaction signals. The following discussion will be limited to non-verbal interaction signals since they are an important part of natural interaction.Non-verbal interaction signals are complex and can often only be detected by analyzing changes over time. An example for such a signal is nodding which cannot be detected in a single picture but a video stream reveals the movement. Therefore a sensor system with abstraction layer and time-dependent analysis is required.

The perception system of the humanoid robot ROMAN realizes the described requirements. Basically it 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 in the sensor data and combines them to a model of the environment. The memory module stores these models over time to allow a comparison with previous capturing and the extraction groups uses this "time-memory" to extract time dependent information. Fig. 2 shows the three basic parts and the information flow between the modules.

3.1 Information Capturing

The information capturing system uses the data from the physical sensors, divides them into several "virtual sensors" and combines the knowledge 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

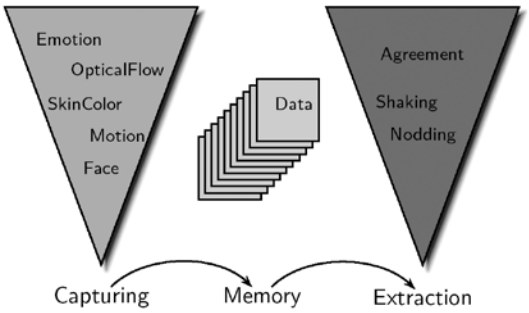

Figure 2: The perception system of the robot with the basic capturing group, the short term memory module and the high-level information extraction group. The data flow is always directed from capturing to extraction.

in respect to a specific property. This concept allows a modular integration of additional sensors for a specific type of information. The perception system for the humanoid robot ROMAN currently has the following virtual sensors:

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.

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. [7]

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

Emotion Detector The emotion detector uses a haar cascade classifier to localize important features in the human face. The feature points are directly mapped to Ekman's action units [4]. Based on the distance of the feature to each other it is possible to calculate the activities of each unit. These activities are analyzed to deduce the emotion of a person in an image.

Color Detector This detector classifies each pixel of the input images and combines connected color regions to color blobs using the CM Vision library. This detector can be used to separate regions with a specific color although it is not suitable for skin color detection. [3]

Motion Detector The motion detector generates difference images of successive pictures to extract regions with high motion activities. This detector does not provide any information about the direction of the motion in an image.

Sound Localizer The sound localization provides 3D information about the most prominent source of sound in the environment of the robot. (work in progress)

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 particle filter on the image plane. The output of the filter and the accoring information from the virtual sensors are combined to a general description of every object in the focus of the robot.

3.2 Information Memory

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 simultaneously tracking of multiple objects. 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 according time stamp.

3.3 Information Extraction

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 activity of each extraction module represents 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.

The activity a of the module nodding is calculated by

$$a = s \cdot \parallel (ampl_y - ampl_x) \cdot 10 \parallel \cdot \parallel ampl_y \parallel \qquad (1)$$

while $\parallel \bullet \parallel$ 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.

4 Experiments

The experiment is performed on a standard PC with an Intel(R) Core(TM)2 Duo CPU E4500 with 2.20GHz. The captured images are scaled to a size of 320x240pixels with a frame rate of 10fps. The perception system with capturing, memory and extraction runs in a cycle time of about 100ms with a CPU load of about 50percent. The particle filter uses 2000 particles.

The experiment was performed in a standard office environment with the standard lightning conditions. A test person has been asked to sit in front of the robot and randomly perform head nodding and shaking actions. During this time the activities of the nodding as well as the shaking behavior has been recorded over time.

Figure 4 show a representative part of the logs during the experiment. The first two graphs show the activity $a \in [0, 1]$ of the nodding and the shaking behavior. A high activity indicates a high probability that the person in the focus of the robot is performing the corresponding action. The third and fourth diagrams show the average optical flow in x and y-direction in pixels. The x-axis in all diagrams indicates the number of the recorded sample.

It can be seen that the person starts nodding at about sample index 3000 leading to a rising activity of the nodding behavior. The activity is high as long as the nodding exists in the ring buffer. The head shaking action starting about sample 19000 does not have any influence on the nodding activity but increases the activity of the head shaking behavior. The response time of both behaviors is mainly influenced by the size of the ring buffer. A large buffer leads to a long time of activity.

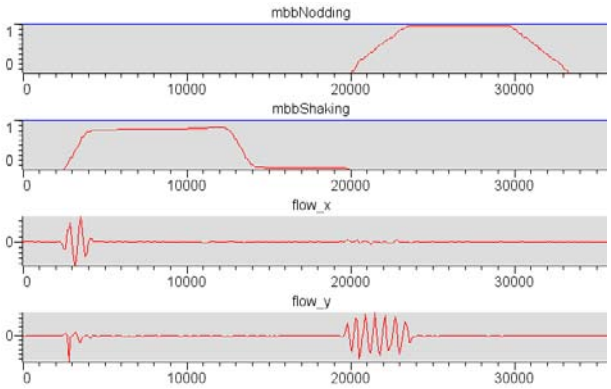

Figure 3: Experimental results of a nodding and head shaking behavior. The x-axis indicates the index of the recorded sample. The first two plots show the activity of the nodding and the shaking behavior. Both behaviors are stimulated during the hole experiment. The third and fourth plot show the average optical flow of the person's face in x- and y-direction in pixels.

5 Summary and Outlook

The paper has introduced a modular sensor system for humanoid robots which is able to extract elementary high level knowledge from high dimensional sensor information. With the help of this concept it is possible to separate the perception system from the rest of the control architecture without limitations in flexibility. The specification of an interface between sensor and control system reduces the amount of transferred data and reduces the complexity of the control system.

The perception system proposed here is subdivided into the three main components capturing, memory and extraction which realizes a modular concept. The capturing uses the physical sensor information to generate basic information like skin color or face position. This information are used as input to a filtering system which generates an abstract view on the observed objects. This abstract information is stored in the memory to reveal time dependent knowledge. The third modules, the extraction uses the memory to generate high level information which are passed to the control system.

A final experiment shows the extraction group with the nodding and head shaking behavior extracted from the information of the average optical flow.

Further developments on the perception system will increase the amount of observable behaviors. To achieve this goal it is necessary to extend all three basic components of

Auszug aus: Künstliche Intelligenz, Heft 1/2009, ISSN 0933-1875, BöttcherIT Verlag, Bremen, **www.kuenstliche-intelligenz.de/order**

the perception system. The capturing system will include audio information in the filtering process and extend the observations to 3D space. The memory system will be extended in a way that long and short term memory are coexisting in a common data structure. And last but not least the extraction system will be extended with more behavior modules and the influence of the emotional state in the extraction system will be added.

References

[1] T. Asfour, K. Regenstein, P. Azad, J. Schröder, and R. Dillmann. Armar-iii: A humanoid platform for perception-action integration. In *HCRS, Second international workshop on Human-Centred Robotic Systems*, Munich, Germany, October 6-7 2006.

[2] P. Azad, A. Ude, T. Asfour, and R. Dillmann. Stereo-based markerless human motion capture for humanoid robot systems. In *IEEE International Conference on Robotics and Automation (ICRA)*, Roma, Italy, April 10-14 2007.

[3] J. Bruce, T. Balch, and M. Veloso. Fast and inexpensive color image segmentation for interactive robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Takamatsu, Japan, October 30 - November 5 2000.

[4] P. Ekman, W.V. Friesen, and J.C. Hager. *Facial Action Coding System (FACS) - Manual*, 2002.

[5] M. Fujita, R. Hasegawa, G. Costa, T. Takagi, J. Yokono, and H. Shimomura. Physically and emotionally grounded symbol acquisition for autonomous robots. In *Proceedings of the AAAI Fall Symposium: Emotional and Intelligent II*, pages 35–36, North Falmouth, USA, November 2001.

[6] 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. to appear.

[7] M.J. Jones and J.M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46((1)):81–96, 2002.

[8] K. Mianowski, N. Schmitz, and K. Berns. Mechatronics of the humanoid robot roman. In *Sixth International Workshop on Robot Motion and Control (RoMoCo)*, Bukowy Dworek, Poland, June 11-13 2007.

[9] F. Schulz von Thun. *Miteinander Reden 1*. Rowohlt Taschenbuch Verlag GmbH, 1997.

Contact
Prof. Dr. Karsten Berns,
Dipl.-Inf. Norbert Schmitz
Robotics Research Laboratory
Department of Computer Science
University of Technology Kaiserslautern
PO Box 3049, 67653 Kaiserslautern, Germany
berns@informatik.unikl.de
nschmitz@informatik.unikl.de

Karsten Berns studied computer science at the University of Kaiserslautern till 1988. In 1994 he received his PhD. from the University of Karlsruhe. From 1989 till April 2003 he was employed at the Research Center of Information Technologies (FZI) at the University of Karlsruhe where he led the group "Interactive Diagnosis- and Service systems" (IDS). Since April 2003 he is a full professor for robotic systems at Kaiserslautern University of Technology. Present research activities are in the area of autonomous mobile robots and humanoid robots with a strong focus on control system architecture and behavior-based control.

Norbert Schmitz studied Informatics at the Technical University of Kaiserslautern till 2005. He wrote his diploma thesis in the area of outdoor robot localization. Since 2006 he is a PhD student in the robots research lab of Prof. Berns focusing on the topic of dynamic modeling of humans for natural human-robot interaction.