International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 3, Issue 5 (December 2013) PP: 24-35

Human-Oriented Interaction with an Anthropomorphic Robot

Bachina harishbabu[1], G. Vijayabaskar[2], Dr. P. N. Chetty[3]
[1]Asst. prof, [3]Asst. prof, [3]Assoc. Prof[3]
Department of automobile engineering,
Vel-Tech Dr.RR & Dr.SR Technical University, Chennai, India

ABSTRACT: *A very important aspect in developing robots capable of human-robot interaction (HRI) is the research in natural, human-like communication, and subsequently, the development of a research platform with multiple HRI capabilities for evaluation. Besides a flexible dialog system and speech understanding, an anthropomorphic appearance has the potential to support intuitive usage and understanding of a robot, e.g .. human-like facial expressions and deictic gestures can as well be produced and also understood by the robot. As a consequence of our effort in creating an anthropomorphic appearance and to come close to a human--human interaction model for a robot, we decided to use human-like sensors, i.e., two cameras and two microphones only, in analogy to human perceptual capabilities too. Despite the challenges resulting from these limits with respect to perception, a robust attention system for tracking and interacting with multiple persons simultaneously in real time is presented. The tracking approach is sufficiently generic to work on robots with varying hardware, as long as stereo audio data and images of a video camera are available. To easily implement different interaction capabilities like deictic gestures, natural adaptive dialogs, and emotion awareness on the robot, we apply a modular integration approach utilizing XML-based data exchange. The paper focuses on our efforts to bring together different interaction concepts and perception capabilities integrated on a humanoid robot to achieve comprehending human-oriented interaction.*

Keywords: *HRI,XML, BARTHOC.*

I. Introduction

For many people, the use of technology has become common in their daily lives, Examples range from DVD players, Computers at workplace or for home entertainment; to refrigerators with integrated Internet connection for ordering new food automatically.The amount of technique in our everyday life is growing continuously. Keeping in mind the increasing mean age of the society and considering the increasing complexity of technology, the introduction of robots that will act as interfaces with technological appliances is a very promising approach. An example for an almost ready-to-sale user-interface robot is iCat [I], which focuses on the target not to adapt ourselves to technology but being able to communicate with technology in a human-like way.

Fig. Interaction with BARTHOC in human-like manner

One condition for a successful application of these robots is the ability to easily interact with them in a very intuitive manner. The user should be able to communicate with the system by, e.g. natural speech, ideally without reading an instruction manual in advance. Both the understanding of robot by its user and the understanding of the user by the robot is very crucial for the acceptance. To foster a better and more intuitive understanding of the system, we propose a strong perception-production relation realized once by human-like sensors as video cameras integrated in the eyes and two mono-microphones, and secondly by a human-like outward appearance and the ability of showing facial expressions (FEs) and gestures as well. Hence, in our scenario, a humanoid robot, which is able to show FEs and uses its arms and hands for deictic gestures interacts with humans (see Fig, I). On the other hand, the robot can also recognize pointing gestures and also coarsely the mood of a human, achieving equivalence in production and perception of different communication modalities. Taking advantage of these communication abilities, the presented system is developed toward an information retrieval system for receptionist tasks, or to serve as an easy-to-use interface to more complex environments like intelligent houses of the future. As a first intermediate step toward these general scenarios, we study the task of introducing the robot to its environment, in particular to objects in its vicinity. This scenario already covers a broad range of communication capabilities. The interaction mainly consists of an initialization by, e,g., greeting the robot and introducing it to new objects that are lying on ,a table, by deictic gestures of a human advisor [2]. For progress in our human-robot interaction (HRI) research, the behavior of human interaction partners demonstrating objects to the robot is compared with the same task in human-human interaction. Regarding the observation that the way humans interact with each other in a teaching task strongly depends on the expectations they have regarding communication partner, the outward appearance of the robot and its interaction behavior are adopted inspired by [3], e.g . adults showing objects or demonstrating actions to children act differently from adults interacting with adults.

For a successful interaction, the basis of our system is constituted by a robust and continuous tracking of possible communication partners and an attention control for interacting, not only with one person like in the object demonstration but with multiple persons enabling us to also deploy the robot for e.g .. receptionist tasks. For human-oriented interaction, not only a natural spoken dialog system is used for input, but also multimodal cues like gazing direction, deictic gestures and the mood of a person is considered. The emotion classification is not based on the content, but on the prosody of human utterances. Since the speech synthesis is not yet capable of modulating its output in reflective prosodic characteristics, the robot uses gestures and FEs to articulate its state in an intuitive understandable manner based on its anthropomorphic appearance. The appearance constraints and our goal to study human-like interactions are supported by the application or sensors comparable to human senses only, such as video cameras resembling human eyes to some extent and stereo microphones resembling the ears. This paper presents a modular system that is capable to find, continuously track, and interact with communication partners in real time with human-equivalent modalities, thus providing a general architecture for different humanoid robots.

The paper is organized as follows. First, we discuss related work on anthropomorphic robots and their interaction capabilities in Section 11. Section III introduces our anthropomorphic robot BARTHOC. In Section IV our approaches to detect people by faces and voices for the use with different robots are shown. Subsequently, the combination of the modules to a working memory-based person selection and tracking is outlined. The different interaction capabilities or the presented robot and their integration are subject to Section V. Finally, experiments and their results showing the robustness of the person tracking and the different kinds of interaction are depicted in Section VI before the paper concludes with a summary in Section VII.

II. Related Work

There are several human-like or humanoid robots used for research in HRI as some examples are depicted in Fig. 2. ROBITA [4], for example, is a torso robot that uses a time and person-dependent attention System to take part in a conversation with several human communication partners. The robot is able to detect the gazing direction and gestures of humans by video cameras integrated in its eyes. The direction of a speaker is determined by microphones. The face detection is based on skin color detection at the same height as the head of the robot. The gazing direction of a communication partner's head is estimated by an eigen face-based classifier. For gesture detection, the positions of hands, shoulders, and elbows are used. ROBITA calculates the postures of different communication partners and predicts who will continue the conversation, and thus, focuses its attention.

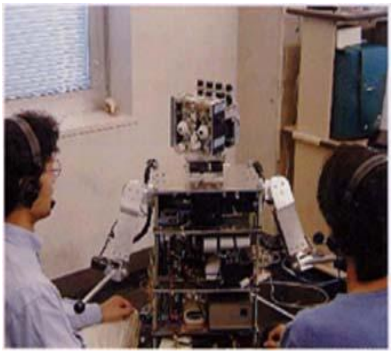


Fig. showing ROBITA

Another robot, consisting of torso and head, is SIG [7], which is also capable of communicating with several persons. Like ROBITA, it estimates the recent communication partner. Faces and voices are sensed by two video cameras and two microphones. The data are applied to detection streams that store records of the sensor data for a short time. Several detection streams might be combined for the representation of a person. The interest a person attracts is based on the state of the detection streams that correspond to this person. A preset behavior control is responsible for the decision taking process how to interact with new communication partners. e.g .. friendly or hostile [7].

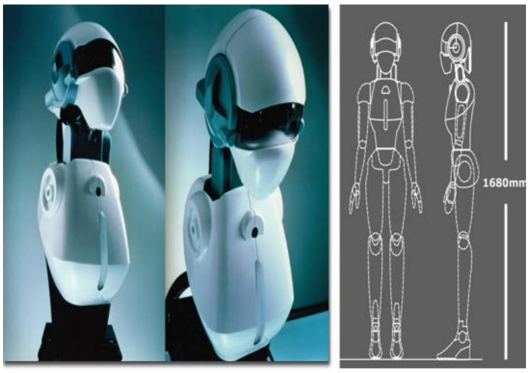


Fig. Showing SIG

Using FEs combined with deictic gestures, Leonardo [8] is able to interact and share joint attention with a human even without speech generation, as prior demonstrated on Kismet with a socially inspired active vision system [9]. It uses 65 actuators for lifelike movements of tile upper body, and needs three stereo video camera systems at different positions combined with several microphones for the perception of its surroundings. Its behavior is based on both planned action, and pure reactive components, like gesture tracking of a person.

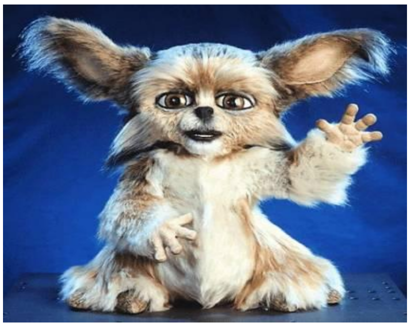


Fig. Showing LEONARDO

An android with a very human-like appearance and humanlike movements is Repliee-Q2. It is designed to look like an Asian woman, and uses 42 actuators for movements from its waist up. It is capable of showing different FEs and uses its arms for gestures, providing a platform for more interdisciplinary research on robotics and cognitive sciences [10]. However, up to now, it has no attention control or dialog system for interacting with several people simultaneously.

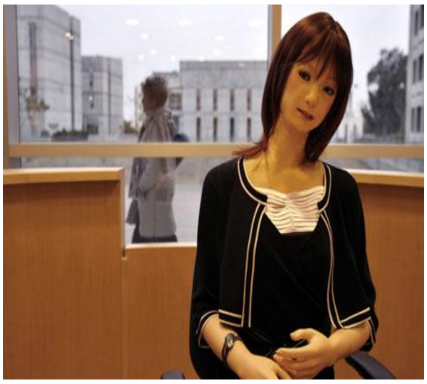


Fig. Showing REPLIEE-Q2

An anthropomorphic walking robot is Fritz [6]. Fritz uses 15 DOFs, 19 to control its body and 16 for FEs. It is able to interact with multiple persons relying on video and sound data only. The video data are obtained by two video cameras fixed in its eyes and used for face detection. Two microphones are used for sound localization tracking only the most dominant sound source. To improve FEs and agility, a new robotic hardware called Robotinho (see Fig. 2) is currently under development.

Despite all the advantages mentioned before, a major problem in current robotics is coping with a very limited sensor field. Our work enables BARTHOC to operate with these handicaps. For people within the sight of the robot, the system is able to track multiple faces simultaneously. To detect people who are out of the field-of-view, the system tracks not only the most intense but also multiple sound sources, subsequently verifying if a sound belongs to a human or is noise only. Furthermore, the system memorizes people who got out of sight by the movement of the robot.

Robot Hardware

We use the humanoid robot BARTHOC [11] (see Fig. 1) for implementing the presented approaches. BARTHOC is able to move its upper body like a sitting human and corresponds to an adult person with a size of 75 cm from its waist upwards. The torso is mounted on a 65-cm-high chair-like socket, which includes the power supply, two serial connections to a desktop computer, and a motor for rotations around its main axis. One interface is used for controlling head and neck actuators, while the second one is connected to all components below the neck. The torso of the robot consists of a metal frame with a transparent cover to protect the inner elements. With in the torso, all necessary electronics for movement are integrated. In total, 41 actuators consisting of dc and servo motors are used to control the robot, which is comparable to "Albert Hubo" [12] considering the torso only. To achieve human-like FEs, 10 DOFs are used in its face to control jaw, mouth angles, eyes, eye brows, and eye lids. The eyes are vertically aligned and horizontally steerable autonomously for spatial fixations, each eye contains a FireWire color video camera with a resolution of 640X480 pixels. Besides FEs and eye movements, the head can be turned, tilted to its sides and slightly shifted forwards and backwards. The set of human-like motion capabilities is completed by two arms. Each arm can be moved like the human model with its joints. With the help of two 5 finger hands, both deictic gestures and simple grips are realizable. The fingers of each hand have only one bending actuator, but are controllable autonomously and made of synthetic material to achieve minimal weight. Besides the neck, two shoulder elements are added that can be lifted to simulate shrugging of the shoulders. For speech understanding and the detection of multiple speaker directions, two microphones are used, one fixed on each shoulder element as a temporary solution. The microphones will be fixed at the ear positions as soon as an improved noise reduction for the head servos is available. By using different latex masks, the appearance of BARTHOC can be changed for different kinds of interaction experiments. Additionally, the torso can be covered with cloth for a less technical appearance, which is not depicted here to present an impression or the whole hardware. For comparative experiments, a second and smaller version of the robot with the appearance of a child exists, presuming that a more childlike-looking robot will cause its human opponent to interact with it more accurately and patiently [13].

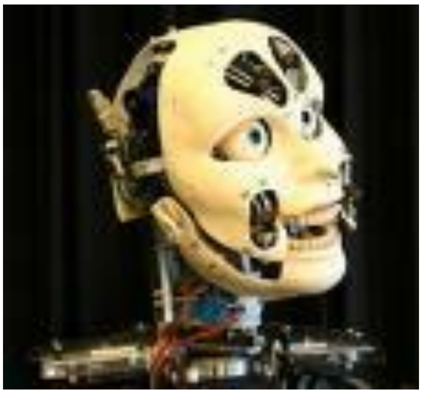
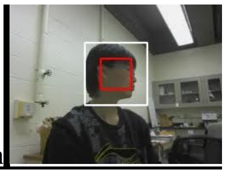


Fig. Showing Robot HardWare

Detecting And Tracking People

In order to allow human-oriented interaction, the robot must be enabled to focus its attention on the human interaction partner. By means of a video camera and a set of two microphones, it is possible to detect and continuously track multiple people in real time with a robustness comparable to systems, using wide field of perception sensors. The developed generic tracking method is designed to flexibly make use of different kinds of perceptual modules running asynchronously at different rates, allowing to easily adopt it for different platforms. The tracking system itself is also running on our service robot BIRON [14], relying on a laser range finder for the detection and tracking of leg pairs as an additional cue. For the human-equivalent perception channels of BARTHOC, only two perceptua1 modules are used: one for face detection and one for the location of various speakers that are described in detail in the following.

A. Face Detection



For face detection, a method originally developed by Viola and Jones for object detection [15] is adopted. Their approach uses a cascade of simple rectangular features that allows a very efficient binary classification of image windows into either the face or non face class. This classification step is executed for different window positions and different scales to scan the complete image for faces. We apply the idea of a classification pyramid [16] starting with very fast but weak classifiers to reject image parts that are certainly no faces. With increasing complexity of classifiers, the number of remaining image parts decreases. The training of the classifiers is based on the AdaBoost algorithm [17]. Combining the weak classifiers iteratively to more stronger ones until the desired level of quality is achieved.

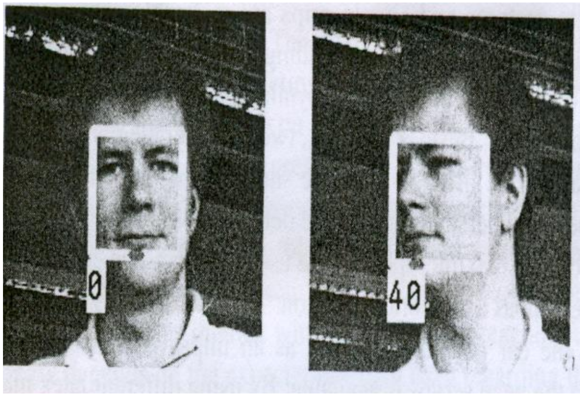


Fig. Showing Facial expressions of a person in different angles

As an extension to the frontal view detection proposed by Viola and Jones, we additionally classify the horizontal gazing direction of faces, as shown in Fig. 4, by using four instances of the classifier pyramids described earlier, trained for faces rotated by 20”, 40", 60", and 80”. For classifying left and right-turned faces, the image is mirrored at its vertical axis, and the same four classifiers are applied again. The gazing direction is evaluated for activating or deactivating the speech processing, since the robot should not react to people talking to each other in front of the robot, but only to communication partners facing the robot. Subsequent to the face detection, a face identification is applied to the detected image region using the eigenface method to compare the detected face with a set of trained faces. For each detected face, the size, center coordinates, horizontal rotation, and results of the face identification are provided at a real-time capable frequency of about 7 Hz on an Athlon64 2 GHz desktop PC with I GB RAM.

B.Voice Detection

As mentioned before, the limited field-of-view of the cameras demands for alternative detect ion and tracking methods. Motivated by human perception, sound location is applied to direct the robot's attention. The integrated speaker localization (SPLOC) realizes both the detection of possible communication partners outside the field-of-view of the camera and the estimation whether a person found by face detection is currently speaking. The program continuously captures the audio data by the two microphones. To estimate the relative direction of one or more sound sources in front of the robot, the direction of sound toward the microphones is considered (see Fig. 5). Dependent on the position of a sound source [Fig. 5(5)] in front of the robot, the run time difference t results from the run times tr and tl of the right and left microphone. SPLOC compares the recorded audio signal of the left [Fig. 5(l)] and the right [Fig. 5(2)] microphone using a fixed number of samples for a cross power spectrum phase (CSP) [19] [Fig. 5(3)] to calculate the temporal shift between the signals. Taking the distance of the microphones dmic and a minimum range of 30 cm to a sound source into account, it is possible to estimate the direction [Fig. 5(4)] of a signal in a 2-D space. For multiple sound source detection, not only the main energy value for the CSP result is taken, but also all values exceeding an adjustable threshold.

In the 3-D space, distance and height of a sound source is needed for an exact detection. This information can be obtained by the face detection when SPLOC is used for checking whether a found person is speaking or not. For coarsely detecting communication partner, outside the field-of-view, standard values are used that are sufficiently accurate to align the camera properly to get the person hypothesis into the field-of-view. The position of a sound source (a speaker mouth) is assumed at a height of 160 Cm for an average adult. The standard distance is adjusted to 110 Cm, as observed during interactions with naive users.

C. Memory-Based Person Tracking

For a continuous tracking of several persons in real time, the anchoring approach developed by Coradeschi and Saffiotti [20] is used. Based on this, we developed a variant that is capable of tracking people in combination with our attention system [21]. In a previous version, no distinction was done between persons who did not generate percepts within the field-of-view and persons who were not detected because they were out of sight.

The system was extended to consider the case that a person might not be found because he might be out of the sensory field of perception. As a result, we successfully avoided the use of far field sensors, e,g .. a laser range finder, as many other robots do. To achieve a robust person tracking two requirements had to be handled. First, a person should not need to start interaction at a certain position in front of the robot. The robot ought to turn itself to a speaking person trying to get him into sight. Second, due to showing objects or interacting with another person, the robot might turn, and as a consequence, the previous potential communication partner could get out of sight. Instead of loosing him, the robot will remember his position and return to it after the end of the interaction with the other person.

It is not very difficult to add a behavior that makes a robot look at any noise in its surrounding that exceeds a certain threshold, but this idea is far too unspecific. At first, noise is filtered in SPLOC by a bandpass filter only accepting sound within the frequencies of human speech that is approximately between 100 and 4500 HZ. But the bandpass filter did not prove to be sufficient for a reliable identification of human speech. A radio or TV might attract the attention of the robot too. We decided to follow the example of human behavior. If a human encounters an unknown sound out of his field-of-view, he will possibly have a short look in the corresponding direction evaluating whether the reason for the sound arises his interest or not. If it does, he might change his attention to it; if not, he will try to ignore it as long as the sound persists. This behavior is realized by means of the voice validation functionality shown in Fig. 6. Since we have no kind of sound classification, except the SPLOC, any sound will be of the same interest for BARTHOC and cause a short turn of its head to the corresponding direction looking for potential communication partners. If the face detection does not find a person there after an adjustable number of trials (lasting. on average, 2 s) although the sound source should be in sight, the sound source is marked as not trustworthy. So, from now on, the robot does not look at it, as long as it persists. Alternatively, a reevaluation of non-trusted sound sources is possible after a given time, but experiments revealed that this is not necessary because the voice validation using faces is working reliable.

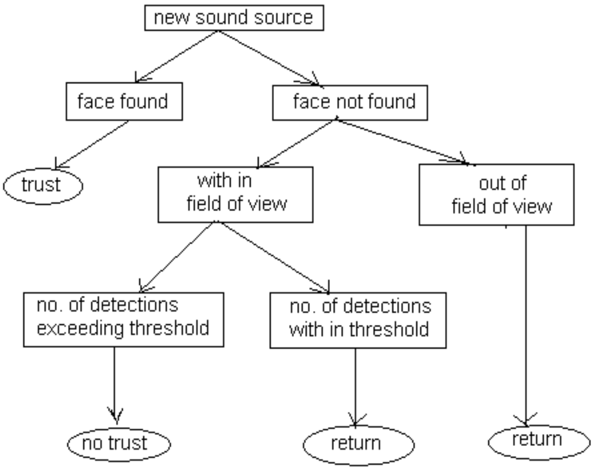


Fig. For Voice Validation

To avoid loosing people that are temporarily not in the field-of-view because of the motion of the robot, a short-time memory for persons was developed. Our former approach [22] loses persons who have nor been actively perceived for 2 s, no matter what the reason was. Using the terminology of [23], the anchor representation of a person changes from grounded if sensor data are assigned, to ungrounded otherwise. If the last known person position is no longer in the field-of-view, it is tested whether the person is trusted due to the voice verification described earlier and whether he was not moving away. Movement is simply detected by comparing the change in a person's position to an adjustable threshold as long as he is in sight. If a person is trusted and not moving, the memory will keep the person's position and return to it later according to the attention system. If someone gets out of sight because he is walking away, the system will not return to the position.

Another exception from the anchoring described in [24] can be observed if a memorized communication partner reenters the field-of-view, because the robot shifts its attention to him. It was necessary to add another temporal threshold of 3 s since the camera needs approximately 1 s to adjust focus and gain to achieve an appropriate image quality for a robust face detection. If a face is detected within the time span, the person remains tracked, otherwise the corresponding person is forgotten and the robot will not look at his direction again. In this case, it is assumed that the person has left while the robot did not pay attention to him.

Besides, a long-time memory was added that is able to realize whether a person who has left the robot and was no longer tracked returns. The long-lime memory also records person specific data like name, person height, and size of some ones head. These data are used to increase the robustness of the tracking system. For example, if the face size of a person is known, the measured face size can be correlated with the known one for calculating the person's distance. Otherwise, standard parameters are used, which me not as exact. The application of the long-time memory is based on the face identification, which is included in the face detection. To avoid misclassification, ten face identification results are accumulated, taking the best match only if the distance to the second best exceeds a given threshold.

After successful identification, person-specific data are retrieved from the long-term memory, if available. Missing values are replaced by the mean of 30 measurements of the corresponding data to consider possible variations in the measurements. Both memory systems are integrated in the tracking and attention module that is running in parallel to the face detection on the same desktop PC at a frequency of approximately 20 Hz. This enables real-time tracking for the interaction abilities with no principal and algorithmic limitations in terms of the number of tracked persons. However, usually, only up to three persons are simultaneous in the field-of-view of the robot, but more persons can be detected by SPLOC, and thus, also robustly tracked for the interaction abilities. On the mobile robot BIRON with a laser range finder, up to ten persons were simultaneously tracked in real time on the basis of continuous perception.

Integrating Interaction Capabilities

When interacting with humanoid robots, humans expect certain capabilities implicitly induced by the appearance of the robot. Consequently, interaction should not be restricted to verbal communication although this modality appears to be one of the most crucia1 ones. Concerning human-human communication, we intuitively read on all these communication channels and also emit information about us. Watzlawick et al. stated their first axiom on their communication theory in 1967 [25] that humans cannot communicate. But, in fact, communication modalities like gesture, FEs, and gaze direction are relevant for a believable humanoid robot, both on the productive as well as on the perceptive side. For developing a robot platform with human-oriented interaction capabilities, it is obvious that the robot has to support such multimodal communication channels too. Accordingly, the approaches presented in the following, besides verbal perception and articulation, also implement capabilities for emotional reception and feedback, and body language with detection and generation of deictic gestures. As communication is seen as a transfer of knowledge, the current interaction capabilities are incorporated in a scenario where the user guides the robot to pay attention to objects in their shared surrounding. Additionally, a new approach in robotics adopted from the linguistic research enables the robot to dynamically learn and detect different topics a person is currently talking about using the bunch of multi modal communication channels from deictic gestures to speech.

A. Speech and Dialog

As a basis for verbal perception, we built upon incremental speech recognition (ISR) system [26] as a state-of-the-art solution. On top of this statistical acoustic recognition module, an automatic speech understanding (ASU) [27] is implemented, on the one hand, to improve the speech recognition, and on the other hand, to decompose the input into structured communicative takes. The ASU uses lexical, grammatical, and semantic information to build not only syntactically but also semantically correct sentences from the results of the ISR. Thus, ASU is capable not only of rating the classified user utterances to full, partial, or no understanding but also of correcting parts of an utterance that seemed to be misclassified or grammatically incorrect according to the information given from the ISR, the used grammar, and the according lexicon. With this information, grammatically incorrect or partially incomplete user utterances are semantically full understood by the system in 84%, in contrast to pure syntactic classification only accepting 68% of the utterances. The output of the ASU is used by the dialog system [28] to transfer the wide range of naturally spoken language into concrete and limited commands to the described robot platform. An important advantage of the dialog management system is it capability not only to transfer speech into commands, but to formulate human-oriented answers and also clarifying questions to assist the user in its interaction with the robot. Based on the current system state, the robot will, e.g., explain why it cannot do a certain task now and how the communication partner can force the robot to complete this task. This verbosity, and thus, the communication behavior of the robot adapts during the interaction with a user, based on the verbosity the interaction partner demonstrates. A more verbose communication partner, e.g., will cause the dialog to give more detailed information on the current system state and to make suggestions which action to take next. A person not reacting to suggestions will cause the dialog to stop with this information. Additionally, the use of more social and polite content by a human-like "thanks" or "please" will cause the same behavior by the robot. Using the initial verbosity in our scenario, where a person shows objects to the robot, the dialog will politely ask, if the system missed a part of the object description or will clarify a situation where the same object gets different labels. The reason might be that there exist two labels for one object or the object recognition failed. This behavior is similar to young children who cannot generalize first and assume a unique assignment between objects and labels, and demonstrates how the described platform is also usable for interdisciplinary research in education between parents and children in psychology.

B. Emotions and Facial Expressions

Not only the content of user utterances is used for information exchange, a special attention has to be paid to the prosody as a meta information. In every day life, prosody is often used, e.g., in ironic remarks. If humans would only rely on the content, misunderstandings would be standard. Thus, it is necessary that a robotic interface for HRI has to understand such meta information in speech. We developed and integrated a software to classify the prosody of an utterance [29] independently from the content in emotional states of the speaker. The classification is trained up to seven emotional states: happiness, anger, fear, sadness, surprise, disgust, and boredom. Now, the robot can, e.g., realize when a communication partner is getting angry and can react differently by excusing and showing a calming FE on its face. The FE of the robot are generated by the facial expression control interface. Based on the studies of Ekman and Friesen [30] and Ekman [3], FE can be divided into six basic emotions: happiness, sadness, anger, fear, surprise, and disgust. Due to constrains given by the mask and user studies, the FE disgust was replaced by the non emotional expression thinking, because disgust and anger were very often mixed up and it proved necessary to demonstrate whether the system needs more time for performing a task by generating a thinking face. The implemented model based on Ekman also considers, besides the current emotional input, a kind of underlying mood that is slightly modified with every classification result. The appropriate FE can be invoked from different modules of the overall system, e.g., BARTHOC starts smiling when it is greeted by a human and "stares" onto an object presented to it. The parameters, basic emotion, emotion intensity, the affect on the mood, and values for a smooth animation can be tested easily by a graphical user interface (see Fig. 9), creating intuitive understandable feedback.

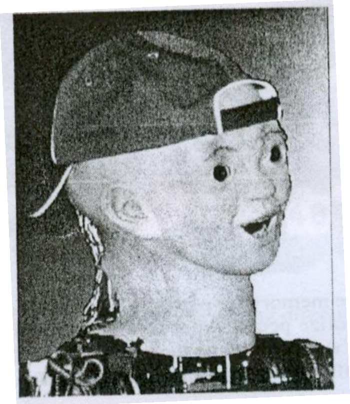


Fig. BARTHOC junior showing happiness

C. Using Deictic Gestures

Intuitive interaction is additionally supported by gestures that are often used in communication [32] without even being noticed. As it is desirable to implement human-oriented body language on the presented research platform, we already implemented a 3-D body tracker [33] based on 2-D video data and a 3-D body model to compensate the missing depth information from the video data. Using the trajectories of the body extremities, a gesture classification [2] is used for estimating deictic gestures and the position a person is referring to. This position does not necessarily need to be the end position of the trajectory, as humans usually point into the direction of an object. From the direction and speed of the trajectory, the target point is calculated even if it is away from the end position. In return, the presented robot is able to perform pointing gestures to presented objects itself, thus achieving joint attention easily (see Fig 10). The coordinates for the tool center point of the robot arm are calculated by the Object Attention and sent to a separate module for gesture generation, GenGes. The module computes a trajectory for the arms consisting of ten intermediate positions for each actuator with given constrains to generate a less robotic-like looking motion instead of a linear motion from start to stop position. The constrains in parallel prohibit the system from being stuck in singularities.

D. Object Attention

As mentioned, deictic gestures are applied for the object attention system (OAS) [2]. Receiving the according command via the dialog, e,g., "This is an orange cup" in combination with a deictic gesture, the robot's attention switches from its communication partner to the referred object. The system aligns its head, and thus, the eye cameras toward the direction of the human's deictic gesture. Using new information from the command like object color, size, or shape, different image processing filters are applied to the video input, finally detecting the referred object and storing both an image and, if available, its sound to a database.

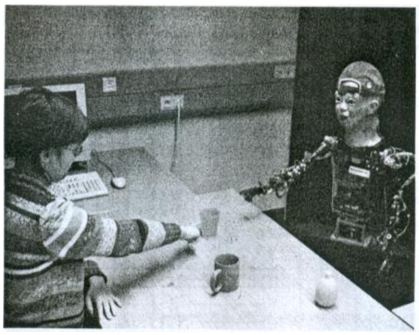


Fig. Using deictic gestures to show and teach BARTHOC junior new objects

E. Dynamic Topic Tracking

For the enhanced communication concerning different objects, we present a topic-tracking approach , which is capable to classify the current utterances, and thus, object descriptions into topics. These topics are not defined previously, neither in number nor named, but are dynamically build based on symbol co-occurrences symbols that co-occur often in the same contexts (i.e .. segments, see paragraph about segmentation later) can be considered as bearing the same topic, while symbols seldom co-occurring probably belong to different topics. This way, topics can be built dynamically by clustering symbols into sets, and subsequently, tracked using the five following steps.

In the preprocessing, first, words indicating no topic. i.e. function words, are deleted. Subsequently, they are reduced to their stems or lemmas. And third, using situated information for splitting up words used for the same object but expressing different topics. The stream of user utterances is ideally separated into single topic during the segmentation step. In our system, segmentation is done heuristically and based on multimodal cues like the information from the face detection, when a person suddenly changes the gaze to a new direction far apart from the first one, or when for a longer period of time it is paused with a certain interaction. Words co-occurring in many segments can be assumed to bear the same topic. Based on co-occurrences, a so-called semantic space is build in the semantic association phase. Semantic spaces are vector spaces in which a vector represents a single word and the distances the thematic similarities. By using distance information, clustering into thematic areas that ideally bear a single topic is applied to the semantic spaces. The tracking finally compares the distances of a spoken word to the centers of the topic clusters and a history, which is created by biasing the system toward the last detected topic if no topic change is detected between the current and the last user utterance.

F. Bringing It All Together

The framework is based on a three-layer hybrid control architecture (see Fig. 11) using reactive components like person tracking for fast reaction on changes in the surrounding and deliberative components like the dialog with controlling abilities. The intermediate layer consists of knowledge bases and the so-called execution supervisor [35]. This module receives and forwards information, using four predefined XML-structures:
1) event, information sent by a module to ESV:
2) order, new configuration data for reactive components (e.g .. actuator interface):
3) status, current system information for deliberative modules
4) reject, current event cannot be processed (e.g .. robot is not in an appropriate state for execution).

Any module supporting this interface can easily be added by just modifying the configuration file of the ESV. The ESV additionally synchronizes the elements of the reactive and deliberative layer by its implementation as a finite state machine. For example, multiple modules try to control the robot hardware.

The Person Attention ends commands to turn the head toward a human, while the Object Attention tries to fixate an object. So, for a transition triggered by the dia1og from ESV state speak with person to learn objects the actuator interface is reconfigured to accept commands by the Object Attention and to ignore the Person Attention. The XML data exchange between the modules is realized via XCF [36], an XML-based communication framework. XCF is capable of processing both 1 to n data streaming and remote method invocation using a dispatcher as a name server. An example for this system is given in Fig. 12, which demonstrates the connection or the hardware specific software controllers to the software applications resulting in a robotic research platform already prepared for extended multimodal user interaction and expendable for upcoming research questions.

Experiments

The evaluation of complex integrated interactive systems as the one outlined in this paper is to be considered a rather challenging issue because the human partner must be an inherent part of the studies. In consequence, the number of unbound variables in such studies is enormous. On the one hand, such systems need to be evaluated as a whole in a qualitative manner to assess the general appropriateness according to a given scenario. On the other hand, a more quantitative analysis of specific functionalities is demanded. Following this motivation, three different main scenarios have been set up to conduct experiments in. However, in the first two scenarios outlined in the following, all modules of the integrated system are running in parallel and contribute to the specific behavior. In the third scenario, the speech understanding was not under investigation in favour of evaluating the prosody information only. The experiments observed the ethical guideline of the American Psychological Association (APA). For the first scenario, two people interact with the robot to evaluate its short-time memory and the robustness of our tracking for more than one person. Additionally, the attention system is tested in correctly switching the robot's concentration to the corresponding person but not to disturb a human-human interaction at all. In the second scenario, one person interacts with the robot showing objects to it evaluating gesture recognition (GR) and object attention. The last scenario concerns the emotion recognition of a human and the demonstration of emotions by FEs of the robot. All experiments were done in an office-like surrounding with common lighting conditions.

A. Scenario 1: Multiple Person interaction

The robustness of the system interaction with several persons was tested in an experiment with two possible communication partners and conducted as described in Fig. 11(a)-(d). In the beginning, a person tracked by BARTHOC. A second person out of the visual field tries to get the attention of the robot by saying "BARTHOC look at me" [Fig, 11(b)]. This should cause the robot to turn to the sound source [Fig. 11(c)], validate it as voice, and track the corresponding person. The robot was able to recognize and validate the sound as a human voice immediately in 70% of the 20 trials. The probability increased to 90% if a second call was allowed. The robot was always accepting a recently found person for a continuous tracking after a maximum of three calls.

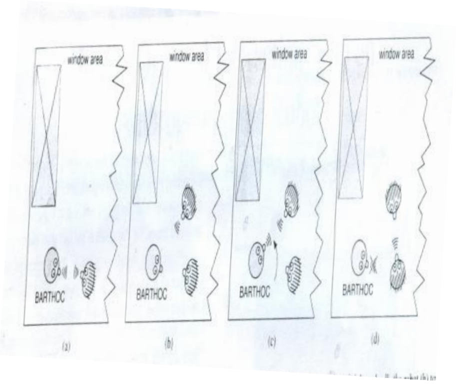


Fig. Person tracking and attention for multiple persons a), a second person stays out of the sight and calls the robot b), to be continuously tracked c) and without losing the first person by use of memory. When two persons talk to each other, robot will focus them but not try to response to their utterances d).

Since the distance of both persons was large enough to prevent them from being in the sight of the robot at the same time, it was possible to test the person memory. Therefore, a basic attention functionality was used that changes from a possible communication partner to another one after a given time interval, if a person did not attempt to interact with the robot. For every attention shift, the person memory used to track and realign with the person temporarily out or view. We observed 20 attention shifts reporting two failures, and thus, a success rate of 90%. The failures were caused by the automatic adoption to the different lightning condition at the person positions that was not fast enough in these two case, to enable a successful face detection for the given time.

Subsequent to the successful tracking of humans, the ability to react in time to the communication request of a person is required for the acceptance of a robot. Reacting in time has to be possible even if the robot cannot observe the whole area in front of it. For the evaluation, two persons who could not be within the field-of-view at the same time were directly looking at the robot, alternately talking to it. In 85% of the 20 trials BARTHOC immediately recognized who wanted to interact with it using the SPLOC and estimation of the gaze direction by the face detector. Thus, the robot turned toward the corresponding person.

Besides the true positive rate of the gazing classification, we wanted to estimate the correct negative rate for verifying the reliability of the gazing classifier too, Therefore, two persons were talking to each other in front or the robot, not looking at it as depicted in Fig. 11(d), Since BARTHOC should not disturb the conversation of people, the robot must not react to speakers who are not look at it. Within 18 trials, the robot disturbed the communication of the persons in front of it only once. This points out that the gaze classification is both sufficiently sensitive and selective for determining the addressee in a communication without video cameras observing the whole scene.

Finally, a person vanishes in an unobserved moment, so the memory must not only be able to keep a person tracked that is temporarily out of the sensory field, but has to be fast enough in forgetting people that were gone while the robot was not able to observe them by the video camera. This avoids repeated attention shifts and robot alignments to a "ghost" person, who is no longer present and accomplishes the required abilities to interact with several persons. An error rate of 5% of the 19 trials demonstrates its reliability.

B. Scenario 2: Showing Objects to BARTHOC

Another approach for evaluating the person memory in combination with other interaction modalities was showing objects to the robot as depicted in Fig. 14(a), resulting in a movement of the robot head that brings the person out or the field-of-view. After the object detection the robot returns to the last known person position. The tracking was successful in 10 of 11 trials; the failure results from a misclassification of the the face detector, since a background very close to the person was recognized as a face and assumed as the primary person. Regions close to the memorized position are accepted to cope with slight movements of persons. Additionally, the GR and the OAS have been evaluated. The GR successfully recognized waving and pointing gestures from the remaining movements of a person in 94% cases with a raise positive rate of 15%. After the GR, the OAS focused the position pointed at and stored the information given from the dialog together with a view.

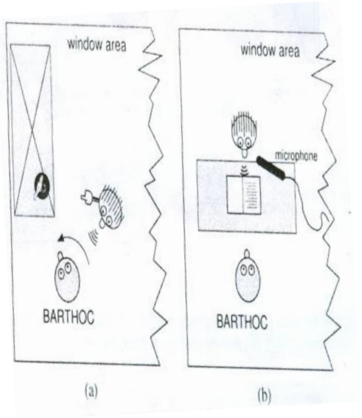


Fig. Interacting with one person. (a) continuous tracking of a person by showing an object to the robot, what causes the communication partner to get out of sight. (b) Person emotionally reads a fairy tale to the robot.

C. Scenario 3: Reading Out a Fairy Tale

For evaluating the emotion recognition and the generation of FEs, we created a setup in which multiple persons were invited to read out a shortened version of the fairy tale Little Red Riding Hood to the robot [see Fig. 14(b)]. For this experiment, the speech processing components were disabled to avoid a distortion of the robots reaction by content of the subject's utterances, in contrast to the previously described scenarios where the overall system was operating. The robot mirrored the classified prosody of the utterances during the reading in an emotion mimicry at the end of any sentence, grouped into happiness, fear. and neutrality. As the neutral expression was also the base expression, a short head movement toward the reader was generated as a feedback for non emotional classified utterances. For the experiment, 28 naive participants interacted with the robot from which 11 were chosen as a control group not knowing that independently from their utterances the robot always displayed the neutral short head movement. We wanted to differentiate the test persons reaction between an adequate FE and an affirming head movement only. After the interaction, a questionnaire about the interaction was given to the participants, to rate on three separate 5-point scales ranging from 0 to 4 the degree as to: 1) the FEs overall fit the situation, 2) BARTHOC recognized the emotional aspects of the story, and 3) whether its response same close to a human.

III. Conclusion

In this paper, we presented our approach for an anthropomorphic robot, and its modular and extendable software framework as both result of the basis for research in HRI. We described components for face detection. SPI.OC, a tracking module based on anchoring [23], and extended interaction capabilities not only based on verbal communication but also taking into account body language. All applications, avoid the usage of human--unlike wide-range sensors like laser range finders or omni-directional camera. Therefore, we developed a method to easily validate sound as voices by taking the results of a face detector in account. To avoid loosing people due to the limited area covered by the sensors, a short-time person memory was developed that extends the existing anchoring of people. Furthermore, a long-time memory was added storing person specific data into file, improving the tracking results. Thus, we achieved a robust tracking in real time with human-like sensors only, now building a basis for user studies using the presented robot as an interdisciplinary research platform. The cooperation with linguists and psychologists will discover new exciting aspects in HRI and advance the intuitive usage and new abilities in interacting with robotic interfaces like the presented one.

References

[1] A.J.N.van Breemen,"Animation engine for believable interactive user interface robots

[2] A. Haasch, N. Hoffmann, J.Frintsch,and G.Sagerer, " A multimodel object attention system for mobile robot"

[3] J.Goetz, S.Kiesler, and A. Powers, "Matching robot appearance and behavior to tasks to improve human-robot co-operation".

[4] Y.Matsusaka, T.Tojo,and T.Kobayashi, "Conversation robot participating in group conversation".

[5] D.Matsui, T.Minato, K. F. MacDorman, and H. Ishiguro, "Genereting natural motion in an android by mapping human motion".