Enabling Multimodal Human–Robot Interaction for the Karlsruhe Humanoid Robot

Rainer Stiefelhagen, Hazım Kemal Ekenel, Christian Fügen, Petra Gieselmann, Hartwig Holzapfel, Florian Kraft, Kai Nickel, Michael Voit, and Alex Waibel

***Abstract*—In this paper, we present our work in building technologies for natural multimodal human–robot interaction. We present our systems for spontaneous speech recognition, multimodal dialogue processing, and visual perception of a user, which includes localization, tracking, and identification of the user, recognition of pointing gestures, as well as the recognition of a person's head orientation. Each of the components is described in the paper and experimental results are presented. We also present several experiments on multimodal human–robot interaction, such as interaction using speech and gestures, the automatic determination of the addressee during human–human–robot interaction, as well on interactive learning of dialogue strategies. The work and the components presented here constitute the core building blocks for audiovisual perception of humans and multimodal human–robot interaction used for the humanoid robot developed within the German research project (Sonderforschungsbereich) on humanoid cooperative robots.**

***Index Terms*—Audiovisual perception, human-centered robotics, human–robot interaction, multimodal interaction.**

I. Introduction

OVER the last decade, much research effort has been focused on the development of humanoid robots, and great progress has been made in developing robots with human-like torsi, including legs, arms, head etc., as well as some human-like motoric skills, such as walking, grasping, dancing [3], [4], [29], [33], [41], [46]. Other researchers have focussed on advancing robots' capabilities in perceiving, interacting, and cooperating with humans [2], [9], [12], [18].

Manuscript received October 14, 2006; revised May 23, 2007. This paper was recommended for publication by Associate Editor C. Laschi and Editor H. Arai upon evaluation of the reviewers' comments. This work was supported in part by the German Research Foundation (DFG) under Sonderforschungsbereich SFB 588—Humanoid Robots. This paper was presented in part at the 13th European Signal Processing Conference, Antalya, Turkey, 2005, in part at the ICSLP, Jeju-Islands, Korea, 2004, in part at the International Conference on Multimodal Interfaces (ICMI), State College, 2004, in part at the KI 2006, Bremen, Germany, in part at the INTERSPEECH, Pittsburgh, PA, 2006, in part at the Proceedings of the 7th International Conference on Multimodal Interfaces, Trento, Italy, October 4–6, 2005, in part at the Sixth International Conference on Face and Gesture Recognition—FG 2004, May, Seoul, Korea, and in part at the First International CLEAR Evaluation Workshop, Southampton, U.K., April 2006.

R. Stiefelhagen, H. K. Ekenel, C. Fügen, H. Holzapfel, F. Kraft, K. Nickel, and M. Voit are with the Interactive Systems Laboratories, Universität Karlsruhe (TH), 76131 Karlsruhe, Germany (e-mail: stiefel@ira.uka.de; ekenel@ira.uka.de; fuegen@ira.uka.de; hartwig@ira.uka.de; fkraft@ira.uka.de; nickel@ira.uka.de; voit@ira.uka.de).

P. Gieselmann is with Lucy Software and Services GmbH, 81643 Muenchen, Germany (e-mail: petra.gieselmann@lucysoftware.com).

A. Waibel is with Carnegie Mellon University, Pittsburgh, PA 15213 USA, and also with the Interactive Systems Laboratories, Universität Karlsruhe (TH), 76131 Karlsruhe, Germany (e-mail: ahw@cs.cmu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TRO.2007.907484

In the framework of the German research center SFB 588 "Humanoid Robots—Learning and cooperating multimodal robots" [1], an interdisciplinary researcher team is working on the development of humanoid robots that can safely coexist, cooperate, and interact with humans in their daily environment. To this end, we focus on the development of an appropriate human-like robotic platform [6] as well as on the development of components that are necessary to facilitate human-friendly, natural human–robot interaction.

Our own research has been the development of such components to enable the perception of the user(s), including many of their important communication cues, such as speech, gestures, head orientation among others, to develop mechanisms to fuse and understand the perceptional and communicative cues, and to build multimodal dialogue components that enable the robot to engage in task-oriented dialogue with their users.

In this paper, we present many of the core perceptual and interaction components that we have developed for the humanoid robots. These include speech recognition, multimodal dialogue processing, visual detection, and tracking and identification of users, including head-pose estimation and pointing gesture recognition. All components have been integrated on a mobile robot platform and can be used for real-time multimodal interaction with a robot. We also report on several human–robot interaction experiments that we conducted. These include experiments on interaction using speech and gestures, the automatic determination of the addressee in human–human–robot interaction, as well as interactive learning of efficient dialogue strategies.

The remainder of this work is organized as follows: In Section II, we give a system overview of the developed components and on information flow between the components. We also give some background on the robotic platform. In Section III, we describe the components for speech recognition, visual perception of the user, and dialogue processing. In Section IV, we, then, present some multimodal human–robot interaction experiments. We conclude the paper in Section V.

II. System Overview

Fig. 1 shows the humanoid robot ARMAR III and its predecessor ARMAR II that are being developed in Karlsruhe within the SFB 588 "Humanoid Robots." For a detailed description of the ARMAR platform, we refer to [6]. The sensors available on the robot head are a stereo camera system and six omnidirectional microphones. For speech recognition, we, alternatively, use a remote close-talking microphone. The current version of the robot has a 1.6-GHz industrial personal computer (IPC) that

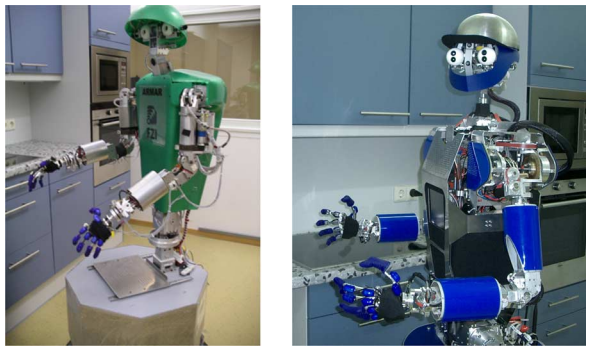


Fig. 1. Humanoid robots (left) ARMAR II and (right) ARMAR III.

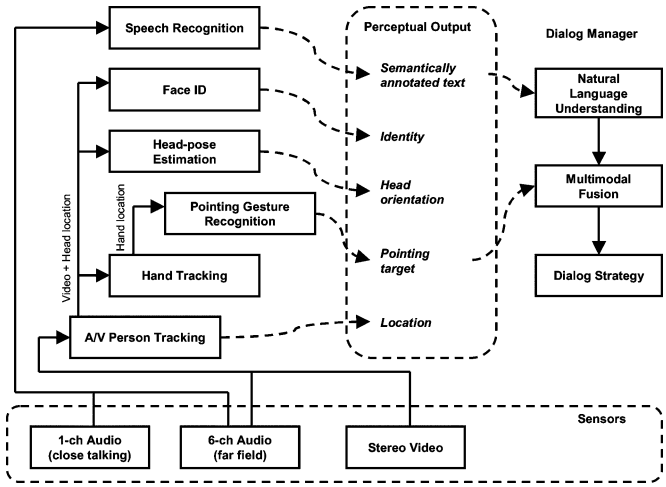


Fig. 2. Overview of the perceptual components.

is reserved for real-time speech recognition, speech synthesis, and acoustic localization. Another 2-GHz PC/104 is to be used for image processing.

Fig. 2 shows an overview of the components employed for human sensing and human–robot interaction. The microphone data is handled by a speech recognition system that produces a transcript of the user's speech. The onboard camera and microphone data are jointly processed by an audiovisual person tracker that fuses both modalities in a probabilistic framework. The output of the person tracker is, then, passed on to the following components: a face identification module based on local discrete cosine transform (DCT) features, a head-pose estimator using artificial neural networks (ANNs), and a 3-D-blob tracker for hand tracking. The hand positions are, in turn, processed by an hidden Markov model (HMM)-based pointing gesture recognizer.

The pieces of information gathered by this perception chain are: semantically annotated speech, the location of the user, his identity and head orientation, and the pointing direction in case of a detected gesture. Natural language understanding, multimodal fusion, and dialogue management operate on the highest level and process semantic information that is delivered by the perceptual components. The following sections describe each of the individual components in more detail.

III. Perceptual Components

A. Speech Recognition

The probably most prominent interaction modality of humans is their speech. In order to provide for natural human–computer interaction, recognition and understanding of spontaneous speech is of utmost importance.

We use the Janus Recognition Toolkit (JRTk) featuring the Ibis decoder [50] developed at Universität Karlsruhe (TH) and Carnegie Mellon University, Pittsburgh for acoustic model training and speech recognition. As linguistic knowledge sources, Ibis is able to use standard statistical n-gram language models or recursive transition networks (RTNs) represented by context-free grammars (CFGs). Using this toolkit, we have developed a user-independent speech recognizer for spontaneous human–robot interaction.

This system is a slimmed down version of the National Institute of Standards and Technology (NIST) Rich Transcription Spring 2004 Meeting Recognition Evaluation (RT-04S) system [37]. The decoding was reduced to only one decoding pass with an already speaker-adaptive trained acoustic model and the search space was further restricted by using tighter beams to reach real-time performance. Incremental speaker adaptation is performed during decoding to increase the recognition performance by compensating different channel and speaker effects. The language model (LM) and vocabulary were tuned toward the robot domain.

B. Sound Event Classification

When moving from close-talking to far-field speech recognition, additional knowledge sources may be exploited. The head of ARMAR III provides six microphones (two in the ears, two in the front, and two in the back of the head) to deliver a 3-D acoustic localization, which is useful for future sound classification and noise suppression. Besides speech recognition, the detection of other perceivable sounds, which can be categorized semantically as observations of dangerous situations (e.g., alarms), human (e.g., speech, coughs) and automated activities (e.g., ring tones), or robot-related sounds (e.g., motor) is of utmost importance.

We performed first experiments [32] to account for environmental sound event classification. Using independent component analysis transformed features on multiframe windows instead of standard mel-scaled frequency cepstral coefficients (MFCCs), significant improvements could be achieved. The models were trained on 4166 real-world kitchen sounds recorded with a distant microphone.

Ergodic tristate HMMs with seven frames context resulted in an average per-class classification error of 9.4% (instead of 14.4% using MFCCs) on a test set of 1826 sound events categorized into 21 sound classes. A Bayesian information criterion was used to determine the number of Gaussians per state automatically. When using Gaussian mixture models (GMMs) with an equal number of Gaussians per model, instead, an average per-class classification error of 9.2% (instead of 12.4%, using MFCCs) could be achieved.

Fig. 3. Video feature maps with the 3-box body model superimposed. (a) Camera image, (b) Foreground segmentation. (c) Detector response. (d) Color segmentation (head–torso–leg color model projected to the red–green–blue channel, respectively).

C. Person Localization and Tracking

Localizinga user in the vicinity of the robot is a prerequisite for many other perceptional components including head-pose estimation and face identification. The person tracking module employed in this system is a further development of the tracker presented in [39]. For a more detailed description, we refer to the earlier-mentioned publication.

In our approach, we fuse both audio and video modalities in a joint particle filter framework. Particle filters [26] represent a generally unknown probability density function by a set of random samples associated with individual weights. Here, each particle is a hypothesis about the 3-D position of the user's head centroid. The particles are weighted by means of a visual and an acoustical observation model.

The features used by the visual observation model are foreground segmentation, face and body detection as well as color models. On the audio side, we localize the user by evaluating the time-delay-of-arrival (TDOA) of the user's speech with respect to the microphones on the robot's head.

1) Video Features: The first visual cue is based on foreground segmentation. The foreground map is made up of the absolute differences between the input image and an adaptively learned background model [see Fig. 3(b)]. To evaluate a particle with respect to the foreground map, a person model at the particle's position is projected to the image. The model consists of three cuboids for head, torso, and legs. The foreground score is, then, calculated by accumulating the foreground pixels covered by this model. This can be done efficiently by means of the integral image [54].

In the face detection algorithm proposed by [54] and extended by [35], a variable-size search window is repeatedly shifted over the image, and overlapping detections are combined to a single detection. In the proposed particle filter framework, however, it is not necessary to scan the image exhaustively: the places to search are directly given by the particle set. Thus, the evaluation of a particle takes only a single run of the cascade. Particles passing many stages of the classifier cascade are given high scores, while particles failing earlier are scored low. In addition to face detection, we use a detector trained on upper bodies. The detector hits are incorporated in the particles' scores using the same method as for face detection.

As a third visual cue, we use individual color models for head, torso, and legs. For each body part, a support map is generated by histogram backprojection [see Fig. 3(d)]. Just like for the foreground segmentation cue, the particles are scored by accumulating the support map pixels under the projected 3-box model. The color models are updated after each frame using the final tracking hypothesis.

2) Audio Features: Given a pair of microphones and a speaker position, the speech signal arrives with a certain TDOA depending on the spatial geometry of the setup. To estimate the TDOA, a variety of well-known techniques [11], [42] exist. Perhaps, the most popular method is the *phase transform* (PHAT), which can be expressed as

$$R_{12}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{X_1(e^{j\omega\tau})X_2^*(e^{j\omega\tau})}{|X_1(e^{j\omega\tau})X_2^*(e^{j\omega\tau})|}\, e^{j\omega\tau} d\omega \qquad (1)$$

where $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of the signals of the microphone pair and τ is the TDOA.

In our approach, we interpret the PHAT as a pseudoprobability density function. The acoustic particle scores are, then, given by the PHAT values at the hypothesized time-delay. We integrate the scores from all those microphone pairs that are exposed to direct sound given the particle's hypothesized sound source location.

D. Face Identification

Face recognition is an important building block of a natural human–robot interaction system. A humanoid robot should recognize the people that it has met before and store associated information related to these people to conduct a natural conversation. However, face recognition in uncontrolled environments is still a difficult problem [57].

To provide robust face recognition and to overcome uncontrolled environmental conditions, we are utilizing multiple samples of the same face that are obtained from the video sequence. Our face recognition system analyzes the input face images to locate the eyes and, then, registers the face images according to the eye center coordinates.

A local appearance-based face recognition approach is used to extract feature vectors from each face image. In this feature extraction approach, the input face image is divided into 8×8 pixel blocks, and on each block, a DCT is performed. The most relevant DCT features are extracted using a zig-zag scan pattern and the obtained features are fused either at the feature level or at the decision level for face recognition [16].

After extraction, the feature vectors are compared with training vectors stored in the database using a nearest-neighbor classifier. Each frame's distance scores are normalized with the min–max normalization method [49], and, then, these scores

are fused over the sequence using the sum rule [30]. The obtained highest match score is compared with a threshold value to determine whether the person is known or unknown. If the similarity score is above the threshold, the identity of the person is assigned to that of the closest match. If the similarity match score is below the threshold, the person is classified as unknown.

This approach was extensively tested on publicly available face databases and compared with other well-known face recognition approaches. The experimental results showed that the proposed local appearance based approach performs significantly better than traditional holistic face recognition approaches. Moreover, this approach was tested on the face recognition grand challenge (FRGC) version 1 data set for face verification [14], and a recent version of it was tested on the FRGC version 2 data set for face recognition [17], and it provided better and more stable results than the baseline face recognition system.

On the FRGC version 2 data set, for example, 80.5% correct recognition rate is obtained under uncontrolled conditions. On the same database, the well-known face recognition algorithms, such as eigenfaces [53], Fisherfaces [8], and Bayesian face recognition [38] attain correct recognition rates of 57.3%, 65.6%, and 63.4%, respectively. In these experiments, there are 120 individuals in the database and each individual has ten training and testing images. There is a time gap of approximately 6 months between the capturing time of the training and the test images. The approach is also tested under video-based face recognition evaluations and, again, provided better results [15].

E. Pointing Gesture Detection

The modules for pointing gesture recognition and hand tracking used in this system have been described in detail in [40]. The gesture recognition is based on hand motion. Dedicated HMMs are used to model the begin-, peak-, and retract-phases of typical pointing gestures. They were trained on 3-D hand trajectories of hundreds of sample pointing gestures. Whenever the models fire successively, a gesture is detected and the peak phase is identified. Within the peak phase, the line between the head and the pointing hand is used to estimate the pointing direction. If the positions of potential pointing targets are known, the most likely target can be selected by measuring its deviation from the pointing direction.

The underlying hand tracking algorithm is based on a combination of adaptive skin-color classification and dense disparity maps. The skin-colored pixels are associated with their corresponding disparity values and are, then, spatially clustered. Those clusters are evaluated in order to find an optimal assignment of head and hands to skin-color clusters in each frame. The optimal assignment maximizes three aspects: 1) the match between hypothesis and observation in terms of skin-color pixels; 2) the naturalness of the hypothesized posture; and 3) the smoothness of transition from the preceding to the current frame. After temporal smoothing, the hand trajectories are passed to the gesture detection module described earlier.

In order to evaluate the performance of gesture recognition, we prepared an indoor test scenario with eight different pointing targets. Test persons were to move around within the robot's field-of-view, every now and then showing the robot one of the marked objects by pointing on it. In total, we captured 129 pointing gestures by 12 subjects. The gestures were manually labeled and, then, compared against the gesture detection hypotheses. The detection rate (*recall*) in this experiment was 80%. The *precision* value—indicating the percentage of detections that are actually pointing gestures—was 74%, meaning that the remaining 26% were false positives. Gestures that were often missed are the ones where the subject points to a target on the floor. In these cases, the pointing hand is close to the torso, which is very similar to the natural resting position. False positives, on the other hand, were mostly caused by errors in the hand tracking module.

In a second experiment, we measured the deviation of the estimated pointing direction and the ideal line from the hand to the target. The average error angle in that experiment was 25°, which allowed us to select the correct target (one out of eight) in 90% of the cases.

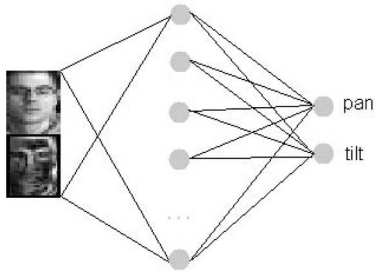


Fig. 4. Neural network for head-pose estimation. As input features, an intensity image of the head region and its Sobel magnitude are used.

F. Head-Pose Estimation

A person's head orientation provides a good indication of a person's focus of attention, i.e., of the objects, area or people which someone is interested in, or with which he or she interacts [10], [36]. This is, in particular, useful to determine whether a person is addressing a robot or someone else [28].

For head-pose estimation, we integrated the basic system that has been described in [55]. We train one neural network classifier to estimate the camera-relative head orientation of a person standing in front of the robot. We crop the head region and rescale it to a normalized size of 64×64 pixels. The normalized intensity image and its Sobel magnitude are merged into a common feature vector. The neural network is trained to output a continuous real-value estimate of the depicted head orientation in the range of $[-90°, +90°]$ in both pan and tilt direction. The network contains 80 hidden units and is trained with standard error backpropagation (100 training cycles). A cross-evaluation set was used to determine the optimal amount of training iterations. Fig. 4 depicts the network's topology and input features.

The core head-pose estimation system was evaluated during the CLEAR'06 evaluation [51], [55]. Here, our approach demonstrated state-of-the-art performance, achieving a mean error of 12.3° for horizontal head-pose estimation and 12.8° for

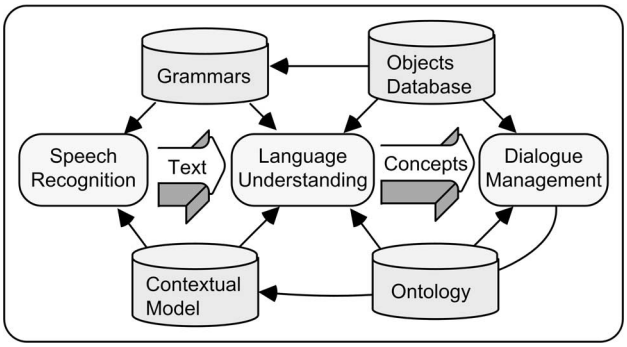


Fig. 5. Dialogue processing framework.

vertical head-pose estimation on the Pointing04 database [21] (see [51] for a comparison to other systems).

G. Dialogue Processing and Fusion

The dialogue manager interprets multimodal input and generates responses and spoken output. Its strategy defines how communication with the user is directed and interpreted. Within the collaborative research center SFB-588, we see the main challenge for dialogue management in providing a natural way to communicate and interact with the robot, and provide the robot with an interface to the environment through which it can obtain new information and learn from communication partners. Natural communication with the robot includes multimodal interaction as well as robust error-tolerant dialogue strategies.

Multimodal fusion as well as dialogue management is performed by the dialogue manager TAPAS [22] that has explicitly been developed for this purpose. It uses the same goal-based dialogue framework as the Ariadne [13] framework. In Ariadne, each dialogue goal is specified by information that has to be collected to achieve this goal. The dialogue strategy selects moves to request missing information or to execute system services. In recent experiments we have used TAPAS to train strategies through reinforcement learning (RL). For this approach the dialog state uses additional information slots that describe an abstraction over information given so far, and actions are selected to optimize the expected reward [44] (see also Section IV-C).

Dialogue management is integrated into the dialogue processing framework as follows (see also Fig. 5): A context-free grammar parses the user utterance. It is enhanced by information from the domain model defining objects, tasks, and properties within the robot's world. The parse tree is converted into a semantic representation, interpreted in context, and added to the current discourse using unification-based update mechanisms. If all information necessary to accomplish a goal is available in discourse, actions associated with the dialogue goal can be executed. Otherwise, clarification questions are generated to acquire missing information. Within the proposed framework, the dialogue manager is tightly integrated with multimodal interpretation and tightly coupled with the speech recognizer by shared knowledge sources (see also Section III-H).

For fusion of speech and pointing gestures we have developed a robust constraint-based approach which is especially suitable to deal with false detections of gestures and improves error-tolerance by incorporating n-best hypothesis lists into the fusion algorithm. This will be further described in Section IV-A.

Since human–robot interaction is prone to communication failures not only due to imperfect speech recognition but also due to user uncertainty about how to interact with such a robot, we developed a general help strategy with all the tasks the robot can accomplish in a hierarchical structure. The user can explicitly ask for help, but there are also implicit factors that suggest that the user is lost and needs help. Within user tests with our robot, we found the following factors indicating problematic situations and the need for help.

1) No speech act can be found in the user utterance.
2) The user utterance is inconsistent with the current discourse (unification with discourse information fails).
3) The user utterance can only partly be parsed.
4) The user utterance is inconsistent with the robot's expectations (unexpected information).
5) The user asks for the same information several times.

All these factors are used for an automatic problem detection: Whenever one of these factors occurs, the user need for help increases and, if it is above a given threshold, the robot explains its capabilities to the user to help him. In addition, the discourse information is finally cleared so that the user can start from scratch again. In this way, we can avoid endless error spirals and support the user in problematic situations.

H. Tight Coupling of Speech and Dialogue Processing

Most current human-machine interfaces consist of three main components: a speech recognizer (ASR), a natural language (NL) parser, and a dialogue manager (DM) [20], [43], [48]. The ASR output in form of n-best lists or lattices is given to the NL parser for interpretation and, then, passed to the DM for making a system decision depending on the current dialogue context.

In [19], we have proposed a tight coupling, which makes a separate NL parser superfluous and the human-machine interface more robust, easier to maintain, and, hence, more portable to new domains. The main structural change is to share the linguistic knowledge sources, i.e., CFGs between IBIS and TAPAS, whereas both components have direct access to the original grammar structure. This gives us the ability to use the results of one system for improving the performance of the other system in the next step by e.g., weighting specific rules in the speech recognizer to direct the decoder into specific regions of the search space. Significant improvements of recognition results, especially on user responses to clarification questions could be achieved. This approach was generalized in [25] by creating an expectation model that describes which utterances are most likely used by the user for his next query/response and gives this information to IBIS. We extend other work, such as [34], by providing a generic approach for semantic grammars that takes into account detailed speech act categorization and information targets [25]. We also use additional knowledge sources, such as missing information in discourse and ontological information in combination with speech act theory, which already exist in the domain description of the dialogue manager.

TABLE I
WER AND SCER FOR THE CONTEXT WEIGHTING APPROACH IN COMPARISON TO THE BASELINE ON ENGLISH (ENG) AND GERMAN (GER) ON TWO SETS: IN-DOMAIN UTTERANCES ONLY AND ALL UTTERANCES

		baseline		context weighting	
		WER	SCER	WER	SCER
ENG	in domain (C)	12.8%	7.8%	10.1%	5.2%
	all (C)	28.3%	15.9%	26.2%	13.7%
	in domain (D)	32.1%	19.9%	29.2%	15.7%
	all (D)	41.9%	26.4%	39.7%	21.7%
GER	in domain (C)	9.8%	4.6%	9.1%	3.9%
	in domain (C)	21.3%	13.1%	20.9%	12.0%

C: close talking speech; D: distant speech.

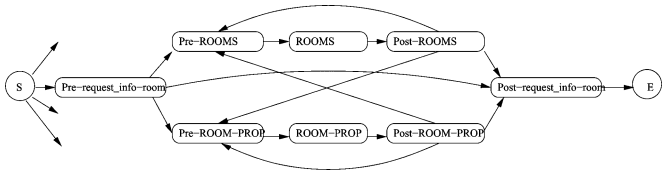


Fig. 6. Example of a schema for the speech act request_information-room. Nodes in rounded boxes are modeled with n-gram LMs, nodes in angled boxes with CFGs.

A recognition system that uses CFGs only as knowledge sources has the disadvantage that out-of-domain sentences or spontaneous user queries are not covered at all. Following the approach presented in [56], we implemented a unified or HMM/CFG language model, which is a combination of statistical n-gram LMs and CFGs. Each speech act is, thereby, represented with the help of a schema modeled as an HMM, whereas each node can be modeled either with a CFG, or with a statistical n-gram LM (see Fig. 6). The training of this model, i.e., the alignment of phrases to specific nodes and the n-gram probabilities, can be done automatically with the help of the EM-algorithm if a sufficient amount of (hand-labeled) data is available. Our experimental results and those presented in [56] show that even on small domains with a limited amount of training data the performance of the unified LM is similar in word error rates to n-gram LMs. Nevertheless, a tight coupling with our dialogue manager is still possible—knowing the structure, i.e., semantical tagged schemas and nodes of the underlying HMM and modeling this by a grammar is sufficient for the dialogue management.

Table I shows a comparison of recognition results[1] for context-dependent weighting of rules on the "barkeeper" scenario, described in Section IV-C. The in-domain set consisting of parseable utterances only contains 267 utterances for English and 152 utterances for German, the full set having 314 utterances for English and 171 utterances for German, respectively. In Table I, it can be seen that the results for the German baseline (without rule weighting) are already very good, probably because the system has been used by people that regularly speak to ASR systems, which is not the case for English. The improvements for the SCER, which is more informative for use by the dialogue system are more significant because it ignores semantically irrelevant errors than the improvement for the WER.

[1]The word error rate (WER) is the average amount of misrecognized words per reference word. The semantic concept error rate (SCER) is measured identical to the WER, but on the level of semantic concepts and their attributes.

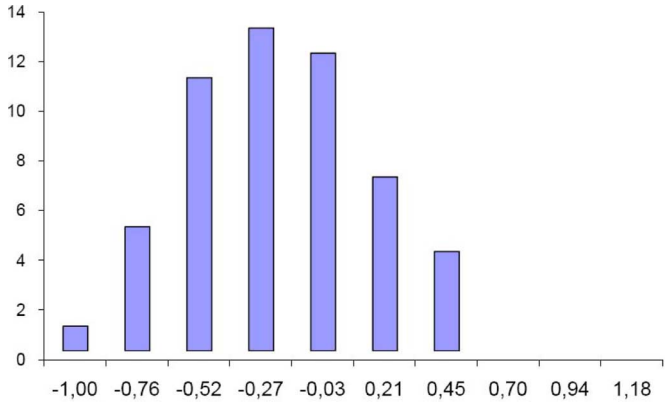


Fig. 7. Time correlation between start of gesture and start of deictic words.

IV. EXPERIMENTS ON MULTIMODAL HUMAN–ROBOT INTERACTION

A. Multimodal Interaction and Fusion

As presented earlier, the system processes multiple modalities, which provide different or redundant information. Some perception outputs are used to provide contextual information about the user such as faceID, others are used for direct interaction such as speech or pointing gestures. Due to different natures multimodal integration happens at different stages. For example, contextual information, such as faceID, is only used in discourse processing, while others can be combined at an earlier stage.

One key aspect of when information from different modalities is combined is their temporal correlation. Speech and deictic gestures, for example, have shown to be highly correlated in time and, thus, are an optimal candidate for fusion after recognition but before dialog processing [23]. Fig. 7 shows this correlation, with statistics of the time difference between a pointing gesture and referring words in speech, such as "this lamp." These statistics emphasize the approach of input fusion instead of fusion in dialog. Furthermore, these statistics were used to define time-based constraints for fusion. Both modalities, speech and gesture, can provide different but useful information, and can also deliver redundant information to improve the robustness of the system. In noisy environments, gesture recognition can improve the results obtained from the speech recognizer, e.g., in bad light conditions, the speech recognizer might deliver information that cannot be obtained through the visual channel. We now describe experiments on fusion of speech and pointing gestures. Later, in Section IV-C, we will also discuss an adaptive learning approach to model information channel that can situationally provide better information.

Our approach to multimodal fusion operates on a pool of input events and uses a constraint-based rule system to determine events that can be merged. Typical constraints take into account the time correlation of the events and their semantic content [23]. The approach is very robust against falsely detected gestures and could be extended to processing of n-best lists. We know of no other approach that takes these aspects into account. Robustness

against falsely detected gestures was an important aspect in our experiments given precision rates of 47% at a gesture detection rate of 87%. This means when achieving high recall of gestures, more than 50% (here 53%) of all detections are false detections, which have to be sorted out by the fusion algorithm. Our approach was validated in an experiment with 500 interaction turns, and 102 multimodal inputs, where the user had to instruct the robot to switch on lamps or bring specific cups. Using pointing gestures has helped to disambiguate objects. The main improvements over a simple fusion algorithm were achieved by robustness against false detections and processing of n-best lists of objects.

B. Determining the Addressee in Human–Human–Robot Interaction

An important issue when developing humanoid robots that should be able to interact with humans is the problem of automatically determining when a robot was addressed by humans and when not. This is an important problem when robots should eventually become companions in our daily lives. A household robot, for example, should know whether a person in the room is talking to him (the robot) or whether this person is talking to someone else in the room.

A person's gaze and head orientation are well known to play an important role during social interaction and to be reliable cues to detect a person's focus of attention [5], [27], [31], [45]. Stiefelhagen *et al.*, for example, have investigated the use of head orientation to detect focus of attention in small meetings [52]. Maglio *et al.* [36] have, for instance, shown that people tend to look toward objects with which they interact by speech. Bakx *et al.* [7] have analyzed facial orientation during multiparty interaction with a multimodal information booth. They found that users were nearly always looking at the screen of the information kiosk when interacting with the system.

In the experiments presented here, we investigate to what extent the addressee of an utterance can be determined in a situation where people interact with each other and a robot intermittently. Our goal is to automatically determine whether the robot has been addressed, and our main approach to accomplish this goal, is to estimate the visual focus of attention of a person during an utterance based on his or her head orientation.

1) Experimental Setup: The data collection setup we used mimics the interaction between two humans and a robot. One person (the *host*) introduces another person (the *guest* to the new household toy, a robot. Our experiments focus on the recordings of the host, since the goal of this work is to determine if the host addresses the robot or the guest. In this experiment, the robot consisted of a construction using a camera to simulate the eyes, and a distant microphone to simulate the robot's ears. We also recorded the host's speech using a close talking microphone. Overall, 18 sessions were recorded, each of roughly 10-min length. The audio data were fully transcribed and tagged on the turn level to indicate whether the host addresses the robot or the guest. For evaluating the video components, we manually labeled the first 2.5 min of the video recordings of four of the sessions.

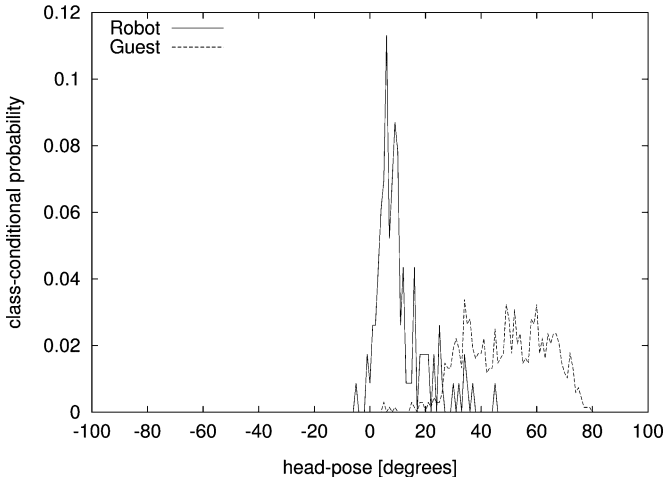


Fig. 8. Typical class-conditional probability distribution for the classification of the visual target for two targets.

2) Relation of Visual Target and the Addressee: To check, how well a person's gaze (i.e., visual focus of attention) can serve as an indicator of the (acoustically) addressed target, we first analyzed the correlation between the manually labeled acoustic targets—i.e., the addressees of an utterance—and the manually labeled visual target, i.e., the targets that had been looked at. Here, the visual targets could be either the "Robot," the other person ("Guest"), or anything else ("Other"). The acoustic targets, as in all our experiments, could be either the "Robot" or the "Guest."

In these experiments, it turned out that looking towards another person is indeed a very reliable indicator that the person was addressed. In fact, in 99.5% of the cases in our data, when the host was looking toward the other person, he indeed addressed that person. Looking at the robot, however, could not be used as such a clear indicator: Here, in only 65% of the cases when a person looked at the robot while talking, the robot also was addressed (for further details, see [28]).

3) Finding the Most Likely Target: Our approach is to estimate a person's visual target based on head orientation and, then, consider this target as the addressee. In order to estimate the visual target, we use a Bayesian approach, where we try to maximize the posterior probability that a certain target has been the visual focus, given the observed head orientation of a person. To compute the *a posteriori* probabilities for the visual focus F_V for each target class T (either "$Robot$" or "$Guest$"), the *a priori* probability $P(F_V = target)$, the class-conditional probability $P(X|F_V = target)$, and the probability $P(X)$ for each horizontal head orientation X have to be estimated

$$P(F_V = T|X) = \frac{P(X|F_V = T) \cdot P(F_V = T)}{P(X)}. \quad (2)$$

When using manually set priors and class-conditional distributions of our model given in (2), we were able to correctly detect the host's visual target in 96% of the frames. Here, the host's head orientation was estimated using the approach described in Section III-F. Fig. 8 depicts typical class-conditional

TABLE II
DETERMINATION OF THE ADDRESSEE, BASED ON ESTIMATING THE VISUAL TARGET. RESULTS WITH TRUE AND LEARNED PARAMETERS

Distribution	Precis.	Recall	F-Meas.	Accu.
True	0.61	0.83	0.7	0.93
Learned	0.6	0.8	0.68	0.89

TABLE III
ACOUSTIC, VISUAL, AND COMBINED ESTIMATION OF THE ADDRESSEE

Estimation	Precis.	Recall	F-Meas.	Accu.
Acoustic	0.19	0.91	0.31	0.49
Head Pose	0.57	0.81	0.67	0.90
Combined	0.65	0.81	0.72	0.92

probability distributions for the two targets, as obtained from the head-pose estimates of one person in our data sets.

Occasions when the person was looking toward the robot could be detected 77% of time, with a relatively high precision of 89%, resulting in an f-measure of 0.82. In another experiment, we, then, automatically learned the priors and class-conditionals of our model in an unsupervised way, as described in [52]. Using these learned model parameters, a slightly lower accuracy and f-measure were obtained. By using these estimated visual targets as an indicator for the addressee of an utterance, we could determine the correct addressee in 93% of the time (89% with learned parameters). Commands toward the robot could be detected with a recall of 0.8 and a precision of 0.6, resulting in an f-measure of 0.7 automatically learned model parameters). Table II summarizes the results.

4) Identification of the Addressee Based on Speech: In addition to using head orientation, we have also investigated the use of acoustic and linguistic features in order to determine the addressee directly from the utterance itself [28]. The features we investigated included low-level speech-based features, such as the length of an utterance and the occurrence of the work "robot" or "robbi" in the utterance, as well as features extracting syntactical and semantical information. We explored classification using various subsets of these features and different classifiers. With the best feature set and a multilayer perceptron for classification, we could detect 91% of the utterances toward the robot (recall), however, with a very low precision of 19% only. By fusing the audio-based determination of the addressee with the head-pose-based approach, a further improvement compared to the vision-only based result could, however, be achieved. Table III summarizes the results.[2]

C. Learning in Human–Robot Dialogues

One of the key aspects of an autonomous system is its ability to learn and to adapt to new environments. We address this issue with two main concepts: first, we implemented and tested an RL approach that allows us to retrain the whole system to adapt to changing environments; second, we have studied techniques that equip the robot with the ability to acquire new knowledge, such as to learn the semantics of words, and to learn to know new objects and unknown persons [24].

[2]Note that these results were computed on different and longer segments than those used to produce the results depicted in Table II (see [28]).

Fig. 9. User interacting with the system in the experimental setup.

Without any doubt, the dialog strategy and its design play a crucial role for the quality of any dialog system. Considering the presence of uncertainty about the user dynamics as well as error-prone system components, the dialog manager's decisions—e.g., which question to ask, initiative to use, or information to confirm—are multifaceted and nontrivial. To overcome the limitations and problematics of handcrafting dialogue strategies, RL approaches have become a promising approach to compute dialog strategies. Since RL algorithms usually require a lot of training data, artificially generated dialogs with a simulated user can be used to handle data sparsity (e.g., [47]).

Here, we present an approach for learning a multimodal dialog strategy by applying a user simulation trained on a small amount of user data, trained for the error conditions of recognition and understanding components. During online operation, more data can, then, be collected and used to refine the models [44].

1) Experimental Setup: For evaluation of the applied methods, we developed a scenario, where the robot acts as an early-stage bartender. Twenty objects, diverging in color, shape, and location are placed on a table in front of the robot and represent the user's order options. Before each dialog, our test subjects silently choose a particular item of interest from the table setting. Our robot, then, initiates a multimodal dialog with the goal to identify the user's object of interest and serve the corresponding item.

Our robot always initiates the dialog with a self-introduction and the mixed initiative action "What can I serve you?". Afterward, the robot has the option to prompt for the three available information slots: object type, object color, and object location. For the case of object location, our robot prompts explicitly for a pointing gesture of the user toward the desired item. Further options exist in confirming information slots either individually or jointly. At any point within the dialog, the robot is able to confirm/exclude the so far best matching item as desired item or, respectively, end the dialog and serve the item. As a general constraint we restricted the number of dialog turns to a maximum of ten.

To be able to conduct the experiments on our experimentation platform, which includes a pan-tilt unit with a mounted stereo camera, distant speech microphones, and close talk microphones, we also equipped the robot with a laser pointer as a pointing device to reference objects on the table. Fig. 9 shows an interaction with the system.

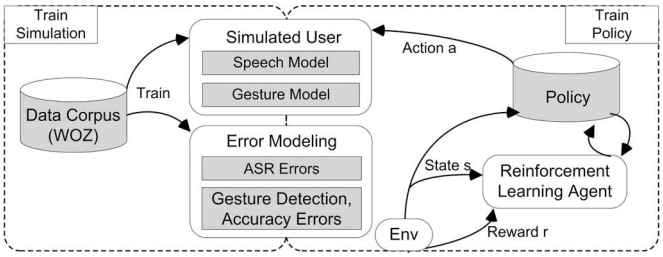


Fig. 10. RL with user simulation and error models.

2) RL: Fig. 10 illustrates the basic components used for training the dialog policy. The user model facilitates simulation of user actions, which is based on a simple bigram model and models the probability of the user uttering a speech act given the previous action by the system $p = P(action_{user}|action_{system})$. Due to its simplicity, it allows us to uniformly model multimodality of user actions, responses to mixed- and system-initiative actions, as well as multiplicity of user actions.

The error model is computed for each component separately and independently. The ASR error model defines common errors by the speech recognizer on the level of the output by the user model, which is a confusion or not-recognition of semantic concepts and in/out-of domain problems. To address interspeaker variations in recognition rates, which might be due to language capability, pronunciation clarity, dialect, or input device adjustment (e.g., proper setup of head-mounted microphone), we stochastically draw from a distribution of error-priors estimated from the existing data corpus. The gesture model was trained on a specific corpus collected for this domain and models detection failures, misdetections, and standard deviations in pointing direction.

Within our Wizard-of-Oz (WOz) experiment, 15 test subjects were engaged and a total of 82 dialogs (314 utterances) were completed. We used this data to train the previously described user and error models, and took subjective user feedback to model subjective criteria for the reward function.

In the system, we, afterward, evaluated the strategy in a different experiment with a new set of users, and also compared the results to a handcrafted baseline strategy. The baseline strategy first collects and jointly confirms the object type and color information, before asking once for a user gesture. Thereafter, the robot explicitly tries to confirm the best matching item within the candidate collection, until the referenced item is confirmed as correct by the user. As a general rule, the dialog is ended as soon as only one candidate item remains.

Eighteen test subjects engaged within this experiment of which only one test subject also participated within the WOz experiment. A total of 94 dialogs (576 utterances) were collected in sequential runs of four to six dialogs of each test subject. Hereby, in order to fairly balance a potential learning effect of the user, we evenly switched between the use of the two strategies. Table IV shows the results of comparing these two strategies The learned strategy performs slightly better than the handcrafted strategy, not only in the simulation (SIM) but also in the real-world (REAL) with respect to the central optimization criterion of collected reward, the task completion rate as well as dialog brevity. As a further positive aspect, despite the small corpus of collected real dialog interactions, the strategies' performance figures resemble nicely for the simulation and real-world domain and, thereby, indicate an adequate accuracy of our simulation model. After retraining the different models with the newly collected data, only 2.7% of the optimal state-actions diverged from the previous run, which indicates that each training cycle converges in a stable manner and that the WOz corpus provides a sufficient amount of training data for such a system.

TABLE IV
COMPARISON BASELINE TO LEARNED STRATEGY

	Baseline Strategy		**RL Strategy**	
	REAL	SIM	REAL	SIM
Dialogs	47	10^5	47	10^5
Task completion	80,4%	83.3%	86.9%	91.3%
Utterances	5.924	5.956	4.943	5.007
Reward	-1.252	-1.067	-0.782	-0.688

V. CONCLUSION

Audiovisual perception of people and their communication modalities, such as speech, gestures, and attention, as well as the fusion and analysis of these cues are prerequisites for building human-friendly robots that can interact with humans.

In this paper, we presented our work on enabling the humanoid robot, which is developed in the German Research Center on Humanoid Robots, with the capability to perceive and interact with people in a natural way. First, we presented several core perceptual components including speech recognition, multimodal dialogue processing, visual detection, tracking and identification of users, as well as head-pose estimation and pointing gesture recognition.

We also reported on several human–robot interaction experiments. These include interaction using speech and gestures, the automatic determination of the addressee in human–human–robot interaction, as well as interactive learning of efficient dialogue strategies.

In the future, we will focus on a seamless integration of these components into an autonomous humanoid robot. Here, further work is necessary in order to improve the efficiency and robustness of all the components. Also, we are investigating how these perception and interaction components can be integrated into an architecture that supports the full range of necessary cognitive capabilities of a humanoid robot, including task planning and supervision, learning, and attention.

ACKNOWLEDGMENT

The authors would like to thank all colleagues in the Interactive Systems Laboratories and in the SFB 588 for their support. In particular, they would like to thank everyone who participated in data collections and user studies.

REFERENCES

[1] The SFB 588 Website. (2004). [Online]. Available: http://www.sfb588.uni-karlsruhe.de/

[2] "Special Issue on Human-Friendly Robots," *J. Robot. Soc. Jpn.*, vol. 16, no. 3, 1998.

[3] K. Akachi, K. Kaneko, N. Kanehira, S. Ota, G. Miyamori, M. Hirata, S. Kajita, and F. Kanehiro, "Development of humanoid robot HRP-3," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, Dec. 2005, pp. 50–55.

[4] R. Ambrose, H. Aldridge, R. Askew, R. Burridge, W. Bluethmann, M. Diftler, C. Lovchik, D. Magruder, and F. Rehnmark, "Robonaut: NASA's space humanoid," *IEEE Intell. Syst. Appl.*, vol. 15, no. 4, pp. 57–63, Jul. 2000.

[5] M. Argyle, *Social Interaction*. London, U.K.: Methuen, 1969.

[6] T. Asfour, K. Regenstein, P. Azad, J. Schröder, A. Bierbaum, N. Vahrenkamp, and R. Dillmann, "ARMAR-III: An integrated humanoid platform for sensory-motor control," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots (Humanoids 2006)*, Genoa, Italy, Dec. 4–6, pp. 169–175.

[7] I. Bakx, K. van Turnhout, and J. Terken, "Facial orientation during multi-party interaction with information kiosks," presented at the Interact 2003, Zurich, Switzerland.

[8] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[9] C. Breazeal, "Social interactions in HRI: The robot view," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 34, no. 2, pp. 181–186, May 2004.

[10] B. Brumitt and J. J. Cadiz, "Let there be light: Comparing interfaces for homes of the future," Microsoft Res., Microsoft Corp., Redmond, WA, Tech. Rep. MSR-TR-2000-92, Sep. 21, 2000.

[11] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 549–557, Nov. 2003.

[12] K. Dautenhahn, M. Walters, S. Woods, K. L. Koay, C. L. Nehaniv, A. Sisbot, R. Alami, and T. Simeon, "How may i serve you?: A robot companion approaching a seated person in a helping context," in *Proc. 1st ACM SIGCHI/SIGART Conf. Human–Robot Interact.* New York: ACM Press, 2006, pp. 172–179.

[13] M. Denecke, "Rapid prototyping for spoken dialogue systems," in *Proc. 19th Int. Conf. Comput. Linguist.*, Taipei, Taiwan, 2002, vol. 1, pp. 1–7.

[14] H. Ekenel and R. Stiefelhagen, "A generic face representation approach for local appearance based face verification," presented at the Workshop Face Recognit. Grand Challenge Exp., 2005.

[15] H. K. Ekenel and A. Pnevmatikakis, "Video-based face recognition evaluation in the CHIL project—run 1," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, 2006, pp. 85–90.

[16] H. K. Ekenel and R. Stiefelhagen, "Local appearance-based face recognition using discrete cosine transform," presented at the 13th Eur. Signal Process. Conf., Antalya, Turkey, 2005.

[17] H. Ekenel and R. Stiefelhagen, "Analysis of local appearance-based face recognition on FRGC 2.0 database," presented at the Face Recogit. Grand Challenge Workshop (FRGC), Arlington, VA, Mar. 2006.

[18] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robot. Auton. Syst.*, vol. 42, pp. 143–166, 2003.

[19] C. Fügen, H. Holzapfel, and A. Waibel, "Tight coupling of speech recognition and dialog management—dialog-context dependent grammar weighting for speech recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, Jeju-Islands, Korea, 2004.

[20] Y. Gao, H. Erdogan, Y. Li, V. Goel, and M. Picheny, "Recent advances in speech recognition system for IBM DARPA communicator," presented at the Eurospeech, Aalborg, Denmark, Sep. 2001.

[21] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial features," presented at the Pointing 2004, ICPR, Int. Workshop Vis. Obs. Deictic Gestures, Cambridge, U.K., 2004.

[22] H. Holzapfel, "Building multilingual spoken dialogue systems," in *Archives of Control Sci., Special Issue on Human Language Technologies as a Challenge for Computer Science and Linguistics (Part II),* vol. 15, no. 4, pp. 555–566, 2005.

[23] H. Holzapfel, K. Nickel, and R. Stiefelhagen, "Implementation and evaluation of a constraint based multimodal fusion system for speech and 3D pointing gestures," presented at the Int. Conf. Multimodal Interfaces, State College, PA, 2004.

[24] H. Holzapfel, T. Schaaf, H. Ekenel, C. Schaa, and A. Waibel, "A robot learns to know people—first contacts of a robot," in *Proc. 29th German Conf. Artif. Intell. (KI 2006), Lect. Notes Artif. Intell.*, Bremen, Germany.

[25] H. Holzapfel and A. Waibel, "A multilingual expectations model for contextual utterances in mixed-initiative spoken dialogue," in *Proc. Interspeech*, Pittsburgh, PA, 2006.

[26] M. Isard and A. Blake, "Condensation—conditional density propagation for visual tracking," *Int. J. Comput. Vis.*, vol. 29, no. 1, pp. 5–28, 1998.

[27] J. W. Tankard, "Effects of eye position on person perception," *Percept. Mot. Skills,* vol. 31, no. 3, pp. 883–893, 1970.

[28] M. Katzenmaier, R. Stiefelhagen, T. Schultz, I. Rogina, and A. Waibel, "Identifying the addressee in human–human–robot interactions based on head pose and speech," presented at the Int. Conf. Multimodal Interfaces, State College, PA, Oct. 2004.

[29] D. Kee, G. Wyeth, A. Hood, and A. Drury, "GuRoo: Autonomous humanoid platform for walking gait research," presented at the Conf. Auton. Minirobots Res. Edutainment, Brisbane, Australia, Feb. 2003.

[30] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[31] C. L. Kleinke, A. A. Bustos, F. B. Meeker, and R. A. Staneski, "Effects of self-attributed and other-attributed gaze in interpersonal evaluations between males and females," *J. Exp. Soc. Psychol.*, no. 9, pp. 154–163, 1973.

[32] F. Kraft, R. Malkin, T. Schaaf, and A. Waibel, "Temporal ICA for classification of acoustic events in a kitchen environment," presented at the Interspeech, Lisbon, Portugal, 2005.

[33] J. Lee, I. Park, J. Kim, and J. Oh, "Mechanical design of humanoid robot platform KHR-3 (Kaist humanoid robot-3: Hubo)," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, Dec. 2005, pp. 321–326.

[34] O. Lemon, "Context-sensitive speech recognition in ISU-dialogue systems: Results for the grammar-switching approach," presented at the Catalog'04, 8th Workshop Semantics Pragmatics Dialogue, Barcelona, Spain.

[35] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in *Proc. Int. Conf. Image Process.*, Sep. 2002, vol. 1, pp. 900–903.

[36] P. P. Maglio, T. Matlock, C. S. Campbell, S. Zhai, and B. A. Smith, "Gaze and speech in attentive user interfaces," in *Proc. Int. Conf. Multimodal Interfaces, Lect. Notes Comput. Sci.* London, U.K.: Springer-Verlag, 2000, vol. 1948, pp. 1–7.

[37] F. Metze, Q. Jin, C. Fügen, Y. Pan, and T. Schultz, "Issues in meeting transcription—The Meeting Transcription System," presented at the Interspeech 2004—ICSLP, Jeju Island, Korea, Oct.

[38] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recogit.*, vol. 33, no. 11, pp. 1771–1782, 2000.

[39] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audio-visual speaker tracking," in *Proc. 7th Int. Conf. Multimodal Interfaces*, Trento, Italy, Oct. 4–6, 2005, pp. 61–68.

[40] K. Nickel and R. Stiefelhagen, "Visual recognition of pointing gestures for human–robot interaction," *Image Vis. Comput.*, to be published.

[41] K. Nishiwaki, T. Sugihara, S. Kagami, F. Kanehiro, M. Inaba, and H. Inoue, "Design and development of research platform for perception–action integration in humanoid robots: H6," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2000, pp. 1559–1564.

[42] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Adelaide, Australia, Apr. 1994, pp. II–273–II–276.

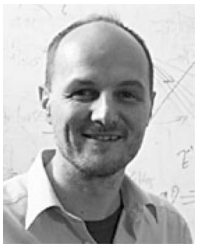
[43] B. Pellom, W. Ward, and S. Pradhan, "The cu communicator: An architecture for dialogue systems," presented at the Int. Conf. Spoken Lang. Process., Beijing, China, Oct. 2000.

[44] T. Prommer, H. Holzapfel, and A. Waibel, "Rapid simulation-driven reinforcement learning of multimodal dialog strategies in human–robot interaction," presented at the Interspeech (Int. Conf. Spoken Lang. Process.), 2006.

[45] J. Ruusuvuori, "Looking means listening: Coordinating displays of engagement in doctor–patient interaction," *Soc. Sci. Med.*, vol. 52, pp. 1093–1108, 2001.

[46] S. Sakagami, T. Watanabe, C. Aoyama, S. Matsunage, N. Higaki, and K. Fujimura, "The intelligent ASIMO: System overview and integration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2002, pp. 2478–2483.

[47] K. Scheffler and S. Young, "Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning," presented at the Hum. Lang. Technol., San Diego, CA, 2002.
[48] S. Seneff, R. Lau, and J. Polifroni, "Organization, communication, and control in the galaxy-II conversational system," presented at the Eurospeech, Budapest, Hungary, Sep. 1999.
[49] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. K. Jain, "Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 450–455, Mar. 2005.
[50] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A one pass-decoder based on polymorphic linguistic context assignment," presented at the Autom. Speech Recognit. Understanding, Madonna di Campiglio, Trento, Italy, Dec. 2001.
[51] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The CLEAR 2006 evaluation," in *Proc. 1st Int. CLEAR Eval. Workshop Multimodal Technol. Percept. Hum., 2006, Lect. Notes Comput. Sci.,* vol. 4122. London, U.K.: Springer-Verlag, 2007, pp. 1–45.
[52] R. Stiefelhagen, "Tracking focus of attention in meetings," in *Proc. Int. Conf. Multimodal Interfaces*. Pittsburgh, PA: IEEE, Oct. 2002, pp. 273–280.
[53] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neursci.,* vol. 3, no. 1, pp. 71–86, 1991.
[54] P. Viola and M. Jones, "Robust real-time object detection," presented at the ICCV Workshop Stat. Comput. Theories Vis., San Diego, CA, Jul. 2001.
[55] M. Voit, K. Nickel, and R. Stiefelhagen, "Neural network-based head pose estimation and multi-view fusion," presented at the 1st Int. CLEAR Eval. Workshop Multimodal Technol. Percept. Hum., 2006, Southampton, U.K.
[56] Y.-Y. Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," presented at the Autom. Speech Recognit. Understanding, St. Thomas, U.S. Virgin Islands, 2003.
[57] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.



Rainer Stiefelhagen received the Diploma degree in computer science in 1996 and the Doctoral degree in engineering sciences in 2002, both from the Universität Karlsruhe (TH), Karlsruhe, Germany.

Currently, he is an Assistant Professor in the Department of Computer Science, Universität Karlsruhe (TH), where he is engaged in research on vision-based perception of humans for human–computer and human–human interfaces. He is a member of the German research center (Sonderforschungsbereich) SFB 588—Humanoid Robots, where he is engaged in research on visual perception of humans and their activities. His current research interests include various aspects of perceptual human–computer interfaces such as audiovisual speech recognition, face recognition, gesture recognition, person-tracking, estimation of head orientations, and modeling the focus of attention of humans.

Hazım Kemal Ekenel received the B.S. and M.S. degrees in electrical and electronic engineering from Boğaziçi University, Bebek, Istanbul, Turkey, in 2001 and 2003, respectively.

He is currently working toward the Ph.D. degree in the Interactive Systems Laboratories, Universität Karlsruhe (TH), Karlsruhe, Germnay. His current research interests include statistical image processing, computer vision, and pattern recognition.

Christian Fügen received the Diploma degree in computer science from the Universität Karlsruhe (TH), Karlsruhe, Germany, in 1999.

Currently, he is a Research Assistant in the Interactive Systems Laboratories, Universität Karlsruhe (TH). His current research interests include the field of automatic speech recognition, especially acoustic and language modeling and adaptation in the context of a simultaneous speech-to-speech translation system.

Petra Gieselmann received the Master's degree from the University of Regensburg, Regensburg, Bavaria, Germany, in 1997, and the Ph.D. degree from the University of Stuttgart, Stuttgart, Germany, in 2007, both in computational linguistics.

From 2002 to 2007, she was a Research Assistant in the Interactive Systems Laboratories, Universität Karlsruhe (TH), Karlsruhe, Germany. Currently, she is with Lucy Software and Services GmbH, Muenchen, Germany, where she is engaged in research on machine translation. Her current research interests include human–robot communication and speech understanding.

Hartwig Holzapfel received the Diploma degree in computer science from the Universität Karlsruhe (TH), Karlsruhe, Germany, in 2003.

Currently, he is a Research Assistant in the Interactive Systems Laboratories, Universität Karlsruhe (TH). His current research interests include multimodal dialogue management and learning.

Mr. Holzapfel is a member of the International Society for Computers and Their Applications (ISCA).

Florian Kraft received the Diploma degree in computer science from the Universität Karlsruhe (TH), Karlsruhe, Germany, in 2005.

Currently, he is a Research Assistant in the Interactive Systems Laboratories, Universität Karlsruhe (TH). His current research interests include sound recognition with applications to automatic speech recognition and environmental sound event classification.

Kai Nickel received the Diploma degree in computer science from the Universität Karlsruhe (TH), Karlsruhe, Germany, in 2003.

Currently, he is a Research Assistant in the Interactive Systems Laboratories, Universität Karlsruhe (TH). His current research interests include person tracking and gesture recognition.

Michael Voit received the Diploma degree in computer science from the Universität Karlsruhe (TH), Karlsruhe, Germany, in 2005.

Currently, he is a Research Assistant in the Interactive Systems Laboratories, Universität Karlsruhe (TH). His current research interests include estimation of head orientation and focus of attention.

Alex Waibel received the B.S. in electrical engineering from Massachusetts Institute of Technology, Cambridge, in 1979, and the M.S. and Ph.D. degrees in computer science from Carnegie Mellon University, Pittsburgh, PA, in 1980 and 1986, respectively.

He is currently a Professor of Computer Science at Carnegie Mellon University, and also Professor at the Universität Karlsruhe (TH), Karlsruhe, Germany. He is associated with InterACT, the International Center for Advanced Communication Technologies at both these universities. At Carnegie Mellon, he also serves as Associate Director of the Language Technologies Institute and holds joint appointments in the Human–Computer Interaction Institute and the Department of Computer Science. He is also associated with the Human Interaction Loop (CHIL) program and the NSF-ITR project STR-DUST. He has founded and cofounded several successful commercial ventures. He was one of the founders of C-STAR, the International Consortium for Speech Translation Research and served as its Chairman from 1998 to 2000. His team has developed the JANUS speech translation system, and, more recently, the first real-time simultaneous speech translation system for lectures. His laboratory has also developed a number of multimodal systems including perceptual meeting rooms, meeting recognizers, meeting browser, and multimodal dialog systems for humanoid robots. He is the holder of several patents. His current research interests include speech recognition, language processing, speech translation, and multimodal and perceptual user interfaces.

Prof. Waibel was the recipient of numerous awards for his work and publications.