A Robot Learns to Know People—First Contacts of a Robot

Hartwig Holzapfel[1], Thomas Schaaf[2], Hazım Kemal Ekenel[1], Christoph Schaa[1], and Alex Waibel[1,2]

[1] InterACT Research, Interactive Systems Labs,
Universität Karlsruhe, Germany
{hartwig,ekenel}@ira.uka.de

[2] InterACT Research, Interactive Systems Labs,
Carnegie Mellon University Pittsburgh, USA
{tschaaf,waibel}@cs.cmu.edu

Abstract. Acquiring knowledge about persons is a key functionality for humanoid robots. In a natural environment, the robot not only interacts with different people who he recognizes and who he knows. He will also have to interact with unknown persons, and by acquiring information about them, the robot can memorize these persons and provide extended personalized services. Today, researchers build systems to recognize a person's face, voice and other features. Most of them depend on pre-collected data. We think that with the given technology it is about time to build a system that collects data autonomously and thus gets to know and learns to recognize persons completely on its own.

This paper describes the integration of different perceptual and dialog components and their individual functionality to build a robot that can contact persons, learns their names, and learns to recognize them in future encounters.

1 Introduction

Recognizing and memorizing other people is an important part of human-human communication. A humanoid robot, if equipped with such functionality, can offer more natural ways of communication and provides a basis to provide personalized services. Today, systems exist that can recognize a person's face, voice or identify persons using other biometric features. In addition, speech recognition and dialog management supply robots with a natural way of communication. We think that with the given technology it is about time to build a system that collects data autonomously and thus gets to know and learns to recognize persons completely on its own. The task of the robot is to meet persons that walk by the system, to establish contact and to find out some information about the person. Using speech recognition, the robot can understand a person's name and learns previously unknown names. The system stores information about these encounters and recognizes the person the next time with a face-ID recognizer. To our knowledge, such a system has not been presented before.

C. Freksa, M. Kohlhase, and K. Schill (Eds.): KI 2006, LNAI 4314, pp. 302–316, 2007.

In this paper we present such a system by describing its architecture, its major components, and first experiments with parts of the system. Techniques that are integrated in the system are speech recognition, unknown word detection, phoneme recognition, spelling, extensible grammars, multimodal dialog management, visual person detection, face detection, face-ID and recognition. We describe the system in two main parts. First, we describe experiments to obtain attention from a person and to maximize success rates in initiating a dialog. Learning mechanisms were applied to classify the intention of the user and to choose actions by the system to obtain attention. We present results of this experiment in the next section. Second, we describe the components and their integration to conduct a dialog with the person during which the system learns the person's name and face-id. The main components that are employed are speech recognition with unknown-word-detection (OOV), face recognition (face-ID) and multimodal dialog management. The components are described and evaluated separately.

The long-term vision of the system is a robot that can patrol the entrance area of a building or a corridor and get in touch with previously unknown people completely on his own. The field of related work is broad due to its interdisciplinary nature. Recently, several robotic systems and humanoid robots that interact with humans have been developed. Our system, as a robot, is broadly related to the field of social robotics [1,2], especially given its style of interaction and the scenario of its employment. An interactive and self controlled robot for example is Lewis [3], a robot with the task to take pictures of groups of persons, visually triggered by certain regions of interest. In our task, the system tracks (moving) persons with whom the robot wants to interact, which relates to visual tracking e.g. [4], but also multimodal tracking e.g. [5] is helps finding the right conversation partner. The following section, where we describe experiments to obtain attention and to initiate a dialog with persons, also relates to studies on engagement [6] in Human-Robot interaction.

Besides controlling a robot that interacts with people, the second great challenge is to conduct a dialog with the person to learn the person's name. Some work exists that describe language learning and learning of words on a multimodal basis. Dusan and Flanagan describe a system for learning words and their meaning in a multimodal setting [7,8]. [9] describes understanding and grounding of new words in situated dialog. Other work focuses on understanding new words, which is a speech recognition task. Chung et al. [10] combine phoneme recognition of spoken input to obtain phonetic representation of names with telephone keypad input to obtain textual representation of names. Even if the name is known but a very large vocabulary list is used, special attention is required. Chung et al. [11] describe a system with a dynamic vocabulary that can be updated according to the given context. The approach from Scharenborg and Seneff [12] runs multiple recognition passes on speech input, with a phone-based OOV word-model in the first step, which is used to constrain the vocabulary in the second step that best matches the resulting phone graph. More related work on unknown words is described in section 3.3.

The paper is organized as follows. Section 2 describes experiments with proactive behavior to obtain the user's attention and initiate a dialog. Section 3 describes the system architecture, the dialog manager, speech recognition and vision. Section 4 concludes the paper and gives an outlook to future work.

2 Obtaining Attention and Initiating Dialogs

A pretest with the system was conducted to evaluate if and how well the robot can obtain the attention of persons passing by, and then initiate a dialog with that person. We furthermore wanted to see which actions of the robot are most important to obtain attention. For the experiment we first recorded and labeled a series of persons passing by the robot, and a series of persons that were told to walk towards the robot. Second, we trained a classifier to classify the person's 'interest' in the system. Finally, we evaluated the success rate of the system to initiate a dialog with interested persons.

For this purpose, we placed the robot in the corridor of our research institute and observed the behavior of people passing by. Each person was then interviewed about the robots behavior and his/her own interest in the system and how much attention was spent to the system when walking by. During each iteration, the robot chose a single action or a combination of the following actions: head movement (turn the head towards the person), play sounds, spoken output. Spoken output was "Hello! Please come closer!" and "Please use the Headset to say Hello!". The success of initiating a dialog was later measured by how many people took the headset to talk to the robot. Figure 1 shows a picture of the robot waiting for persons.

The baseline for attention was estimated by interviewing six persons that have never been in the building before. They were sent to an office at the other end of the corridor and had to pass by the robot, which didn't show any behavior. When they arrived at the office they were interviewed how they had perceived the robot. It was interesting to hear that only one person had even noticed that there was a robot. Afterward, ten more people passed by the robot, in this case the robot reacted by playing a moderate honking sound. Four persons didn't notice the sound, six persons noticed the sound, three of them also interpreted the sound as a reaction of the robot. The baseline can be used to compare the following set of artificial experiments, where persons walked intentionally past the robot to judge its actions with this set of "uninformed" persons. It also shows that the robot itself is not eye-catching at all, and thus the robot's actions become more important to establish a dialog.

In a second experiment, we instructed eleven persons to walk by the robot and judge its actions. Each user had to do this five times, each iteration with different combination of actions. Table 1 shows the results of the experiment. The evaluated categories are 'eye-catching' (does the action influence my attention?) and 'suitable' (do you feel you should start a dialog with the system?). The values for eye-catching are scaled to 0 (no influence), 1 (medium), 2 (annoying). The values for 'suitable' are scaled to -1 (not suitable), 0 (a little bit), 1 (yes). The

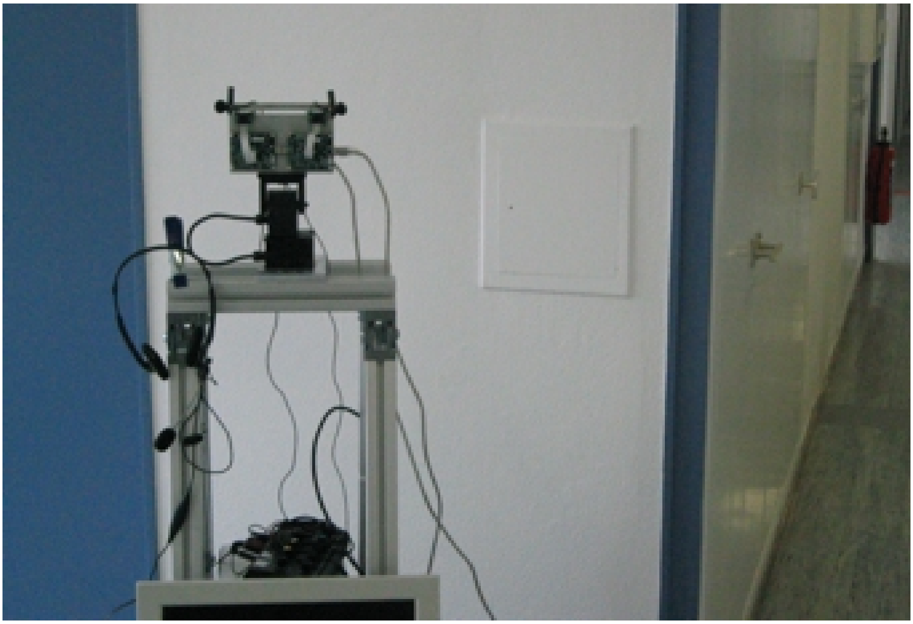

Fig. 1. A picture of the robot, waiting for persons to interact with

Table 1. Evaluation of different system actions

action	eye-catching	suitable
play sound	0.9	-0.3
turn head	0.9	-0.3
say ’hello’	0.9	0.8
play sound then say ’hello’	0.9	0.3
turn head then say ’hello’	1.0	0.9

experiment shows that playing a simple sound is already judged as eye-catching. It should be noted that turning the head also makes some noise induced by the pan-tilt unit. Turning the head offers no increase in attention over playing a simple sound, but leads to the highest attention rate and suitability for initiating a dialog when combined with speech.

The final experiment combines actions to obtain the user’s attention and actions to initiate a dialog with an interest classifier. The classifier was trained only on 3D tracking data from the robot’s stereo camera. This differs from other work that use a laser scanner to determine the position of persons, and vision-based face tracking doesn’t seem to achieve the same reliability. Interest classification thus had to be made robust against recognition and tracking errors. We trained a multi-layer perceptron (MLP) with two hidden layers to classify ’0’ (not interested) or ’1’ (interested). The input features that led to the best results are (i) distance between person and robot (ii) angle between straight robot view (straight ahead) and person (iii) walking speed of person (iv) angle between person's walking direction and robot. The MLP was trained on 1000 samples of the (partly) noisy input data that was provided by the person tracker and hand-labeled interested/not-interested tags. The error rate was on average 14.6% on unseen data, averaged over five runs with different clustering of training, cross-evaluation and test-sets.

The final evaluation then aims to evaluate how well the robot was able to initiate a dialog. It was conducted in 100 attempts distributed over five persons. The experiment was artificial in a way that the persons redid the same experiment a couple of times and decided for themselves if they would interact with the robot. The absolute numbers are thus subjective to the willingness of the persons to interact with the robot. The system chose among four modes. All four modes use different actions to obtain the person's attention, but share the same spoken output once the person is recognized to be 'interested'. In the first and second mode, the system first plays a sound when detecting the person and moves the head towards the person. In the second mode the state transitions don't rely on a single continuous tracking in contrary to the first mode. In the third mode, the system first plays a sound when detecting the person, moves the head towards the person and then follows the person with the head, and also doesn't rely on a single continuous tracking. In the fourth mode, the system only plays a sound when detecting the person, and requires a continuous track again. In all modes, the system says "hello, please come closer!" when the user is interested and then "use the headset to say hello" when the user remains interested to start a dialog.

During some of the iterations the users could not be tracked correctly, e.g. due to changing light conditions. When the system failed to track the user no interaction could be initiated. Table 2 (left columns) shows for each user first the tracker-recognition rate and second the success rate to start a dialog. The table shows a dependency of recognition rate to success rate, but also different behavior by different users. Evaluated on a per system-action modes base 2 (right columns), the results show a smoother distribution of the success rates. It also shows that the modes that were more forgiving regarding the person tracks received higher success rates. Figure 2 shows a series of pictures taken by the robot camera during a single interaction.

Table 2. Evaluation of the success rates per user (left three columns), and evaluation of the success rates per mode (right three columns)

person no.	recognition rate	success rate	mode no.	recognition rate	success rate
1	80%	35%	1	76%	42%
2	100%	85%	2	80%	55%
3	60%	30%	3	76%	58%
4	70%	15%	4	96%	45%
5	100%	50%			

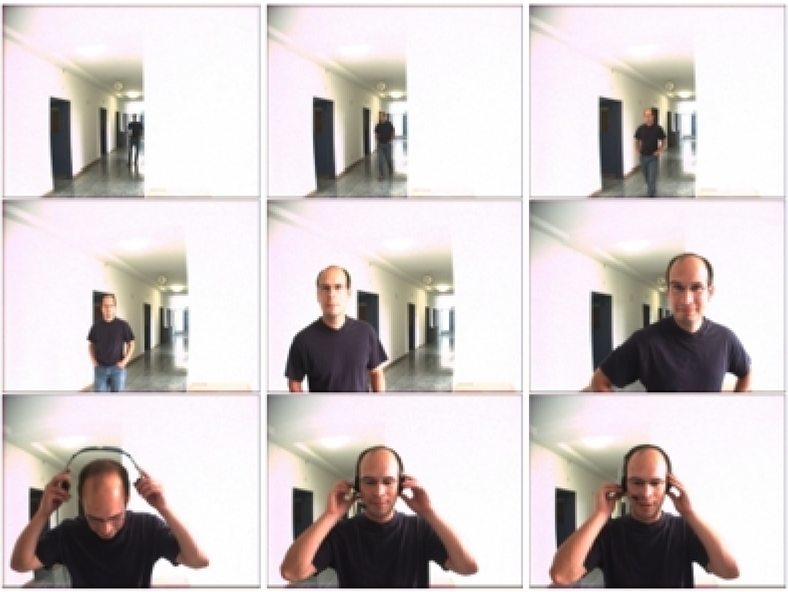

Fig. 2. Series of pictures taken by the robot camera during a single interaction

Tracking of persons was realized with a state-of-the art multi-person tracker developed within the CHIL-project[1]. It uses single 2D-images to detect face-candidates with a haar-cascade based face-detector. 3D-information is obtained for each face-candidate by using disparity information from matching the left and right camera images, including a size estimation for head candidates. More details on the implementation can be found in [13].

3 Recognizing Persons and Names

After initiating a dialog, it is the task of the system to identify the person using snapshots of the person's face and speech input to understand the person's name. To facilitate this task we use a face tracker and keep the person to communicate with in the field of view. A face-ID recognizer computes hypotheses of all snapshots and recognizes either a known or an unknown person. The spoken dialog part is then to ask for the unknown person's name or confirm the name of a known person.

Information about persons, i.e. video snapshots, pictures, voice input, their ID, their names as string and phonetic representation, and information about their latest interactions are stored in a MYSQL database. This database is filled during interaction and is used by the face-ID recognizer to build a model of known persons. It is also used by the speech recognizer and the dialog manager, which read names and their phonetic representations to create grammars for speech recognition, understanding and spoken output.

[1] *http://chil.server.de*

3.1 Face Recognition

Face recognition first tries to determine whether the person in question is known to the system or not. If the person hasn't been stored in the database yet, the robot asks his/her name to enroll the person's face images to the face database labeled with his/her person name.

The face recognition system uses a face sequence, which is provided by a face tracker module, as input. It processes the frames to locate the eyes and then aligns the face images according to the eye center coordinates. The features that will be used for classification purposes are extracted from each face image by using a local appearance-based face recognition approach [14]. In this feature extraction approach, the input face image is divided into 8x8 pixel blocks, and on each block discrete cosine transform (DCT) is performed. The most relevant DCT features are extracted using the zig-zag scan and the obtained features are fused either at the feature level or at the decision level for face recognition [14]. The approach is extensively tested on the publicly available face databases and compared with the other well known face recognition approaches. The experimental results showed that the proposed local appearance based approach performs significantly better than the traditional face recognition approaches. Moreover, this approach is tested on face recognition grand challenge (FRGC) version 1 data set for face verification [15], and a recent version of it is tested on FRGC version 2 data set for face recognition [16]. In both tasks the system provided better and more stable results than the baseline face recognition system. For example, in the conducted experiments on the FRGC version 2 data set, 96.8% correct recognition rate is obtained under controlled conditions and 80.5% correct recognition rate is obtained under uncontrolled conditions. 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 +six months between the capturing time of the training and the test set images. The approach is also tested under video-based face recognition evaluations and again provided better results [17,18]. For details please see [14,15,16,17,18].

After extracting the feature vectors from each face image in the sequence, they are compared with the ones in the database using a nearest neighborhood classifier. Each frame's distance scores are normalized with Min-Max normalization method [19], and then these scores are fused over the sequence using the sum rule [20]. 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 with that of the closest match. If the similarity match score is below the threshold, then the person is classified as unknown. In this case, the robot asks for person's name and saves his/her face images to the database.

3.2 System Integration and Dialog Components

The dialog implementation is based on the Tapas dialog manager [21,22]. It uses a language and domain independent dialogue engine with discourse representation and goal-based dialogue strategies. The dialogue engine follows the Ariadne dialogue manager [23].

In our communication centered scenario, the dialogue manager is the main component to decide which actions to take. Its dialog strategy defines which information to request or which actions the system should take. The dialogue manager also comprises interpretation of multimodal input and is responsible for storing information in the database. All components communicate over a message-based architecture. Figure 3 shows a diagram visualizing the integration of the recognition and understanding components into the dialogue system. The speech recognizer sends an n-best list of parse trees to the NLU-component that converts the parse trees to semantics, formulated as typed feature structures (TFS). The input TFS is converted and interpreted in the dialogue context and finally updates the discourse. Typed feature structures (TFS) also represent referenced database objects and user model data.

The dialogue engine's strategy matches discourse states to dialogue goals, each layer in the discourse corresponds to an unfinished dialog goal.

All semantic concepts that are used to represent user input and discourse representations are defined in an ontology which provides inheritance information and relations between concepts. The dialogue engine is language and domain independent. Language specific parts are semantic grammars for natural language understanding and generation templates for spoken output. Grammar resources for input understanding are an extension to JSGF grammars that are extended with inheritance rules and semantic construction rules. The grammars are shared by the dialogue manager's NLU parts and the speech recognizer.

For speech recognition we use the Janus speech recognizer (JRTK)[24] with the Ibis single-pass decoder[25]. Ibis allows to decode with context-free grammars (CFGs) instead of statistical n-gram language models, and offers a tighter integration of the dialogue manager and Janus by being able to weight grammar rules depending on the dialogue context [26].

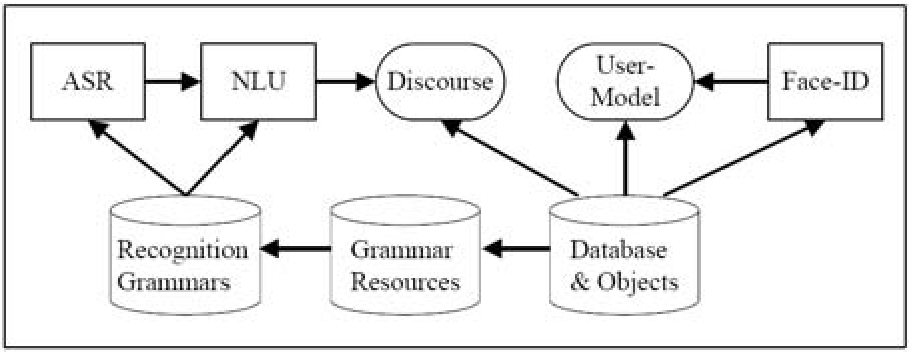

Fig. 3. Flow diagram visualizing the integration of recognition and understanding components, as well as face-ID into the dialogue system

To be able to operate in dynamic environments, i.e. especially to be able to update and understand new names, the system ontology and grammars need to be updated during runtime. In our system, we store person information such as names in an external database. The database information is used by dynamic (grammar) nodes that generate grammar parts from database information, and are update during runtime [27]. These nodes implement a caching mechanism and can be updated from the database when necessary, e.g. when new user names are added.

The following example shows a spoken interaction between the system and a user (table 3). The dialog shows a simple interaction where a person is asked for his name. In case of an unknown name, the system asks a second time to get a better phoneme hypothesis and then asks for the spelling. In the scenario shown here, the dialog manager doesn't try to confirm the spelling. In contrast, the strategy doesn't need to rely on the correct spelling (letters) but uses in the first place a phonetic description, respectively a joint hypothesis as described in section 3.3 that can be used to recognize the name again the next time.

Table 3. A typical dialog with the system

System	Hi, please tell me your name
User	My name is Stephan
System	Can you please repeat this?
User	I Said, my name is Stephan
System	I haven't heard this name before, please spell it.
User	S T E P H A N
System	Thank you. I will recognize you next time.

3.3 OOV Recognition

The above section describes how information about a person and especially the name, can be updated in the system to be recognized with the next input. However, to recognize a new name for the first time (i.e. previously unknown name of a person) a different approach has to be taken. Firstly, the speech recognizer needs to detect that the user has spoken a word which is not covered by the grammar at this point and also not in the vocabulary. This is called unknown-word or out-of-vocabulary (OOV) detection [28,29].

The Hidden Markov Model framework is currently the state of the art in speech recognition. In this framework, the possible output is constrained to the search vocabulary of the recognizer. The search vocabulary usually consists of words with some additional words to cover also acoustic events like hesitations (aem), lip-smack, silence and general noise events. For English we use the simple definition of a word as everything that is written between two blanks. Therefore if a word is spoken that is not in the search vocabulary the recognizer cannot hypothesize it. Actually, the recognizer usual comes up with one or more words that are the closest match to the spoken word. Because the recognizer combines an acoustic model and a language model, in general the words of the closest match do not always sound similar to what was said, or otherwise sounds similar but doesn't fit in the context, or even a combination of both. Hetherington [30] found that in English on average the number of errors a speech recognizer makes per OOV word is larger than one (1.5 - 1.8).

The standard approach to address the Out Of Vocabulary (OOV) problem is to increase the vocabulary size until the average number of OOV words becomes small and the effect on the word error rate becomes small. However, the word error rate does not measure the importance of words and it is obvious that names are very important. If they are missing, it is very hard to understand what is going on. If we would know all names we need in advance this would be nice, but this is in general not the case. Unfortunately, adding all names we know to the search vocabulary can harm the recognition also by increasing the acoustic confusability or because of the low probability of the names, they are still misrecognized. Nevertheless, most important, we just do not know all names that can come up, especially if we think about names that do not exist yet. Therefore we need a dynamic way to extend the vocabulary on demand.

With our approach, the model is extended with special 'words' that represent any unknown sets of words. In a dialog, missing information is acquired and the search vocabulary and language model - here a CFG-grammar - is updated.

Similar to the pioneer work by Asadi [31] who was the first to address the OOV problem in speech recognition, the possible location for OOV words is constrained by a grammar. He investigated two extensions of the recognition model to cover the acoustics of an unknown ship-name, a simple (flat-) acoustic model that is trained on multiple phonemes at the same time and duplicated this model to achieve a minimum length constrained. His findings inspired many researchers to find good acoustic models, and is still the prototype of the current state of the art.

The main goal is to extend the acoustic recognition model with generic words, which give a lower probability compared to the correct word model if the word is known, but are preferred if the word is unknown. In our approach we use Head-Tail-Models (HT-Models) [29] which are a combination of the same precise acoustic models that is used to model the context dependent phonemes of the known words and a generic phoneme model that is trained with multiple phonemes similar to the flat model of Asadi. In [28] this approach was advanced to define the head part of the HT-Models based on the search vocabulary of the recognizer. This has lead to vocabulary optimized HT-Models that also fit very well in a phonetic prefix tree or other search graph structure.

The basic idea of the HT-Models is that it is not possible to decide already after the first phonemes if a word is out of vocabulary because usually the prefix is shared by many known words. Therefore, the head part of the model has the task to compete with the known words until the OOV word starts to diverge from all known words, in which the average likelihood of the exact models gets worse.

If the hypothesis of the speech recognizer indicates the presence of an unknown word, this triggers a dialog with the user to verify and learn the new word.

To add a new word, it has to be included in the search vocabulary and the language model. To extend the search vocabulary, we need a phonetic description of the new word that represents the pronunciation. For the robot domain, this pronunciation usually serves two purposes. One is to allow the pronunciation of the word, e.g. if it is a name, to address a person and therefore it requires a high quality for speech synthesis. The other purpose is to describe the word close enough, that it is not confused with other known words and can be recognized later again. However, for this purpose the chosen phoneme sequence can be less than perfect. There are two sources from which a phoneme sequence can be derived, from a spoken version of the word and from the written form using grapheme to phoneme rules. To get the written form of the name to learn the robot asks the user to spell it. The recognition accuracy for this task is about 90% using a statistical N-Gram model allowing only letters and some human noises. On average, 60% of spelled words have no error, which is comparable to the performance from Hild without vocabulary constrained [32]. Therefore, it is useful to use the top N hypotheses to generate pronunciations. However, generating phoneme sequences for names from graphemes is difficult, especially if the names also come from different languages. In addition, each written form can generate more than one possible pronunciation. So it is necessary to select a small set of pronunciations that is closes to the spoken form. This can be done by using the original utterance to cross-validate. If a reasonable pronunciation was generated by the speech recognizer it replaces the OOV-symbol after the hypothesis is redecoded.

We use the same approach to find a good phoneme sequence, by asking the user to say the word again and do phoneme recognition. The phoneme recognition accuracy is about 65%. Therefore the phoneme recognizer also creates an n-best list of phoneme sequences which are merged with the list generated by the spelling recognizer. Actually, the cross-validation is performed on the merged set. The pronunciation that wins is used to extend the speech recognizer. If the winning pronunciation was generated by grapheme to phoneme, then the attached grapheme sequence is used, otherwise the first best grapheme sequence. If none of the learned pronunciations appear in the redecoded utterance, this indicates that the found phoneme sequences do not generalize well and therefore are not suitable for recognition. This allows to ask the user for more help until the name is learned or to give up learning the name.

3.4 The Dialog Manager

The dialog manager is based on the Tapas dialog framework [21,22] which is described briefly above. Here, we want to describe the extensions that were made to the framework to handle special requirements of the given task. We use most of its existing functionality, such as language understanding, discourse modeling, and knowledge representation, but implemented new dialog strategies that lead in the direction of better social communication. Also the abilities to handle multimodal information [33,27] was extended to handle new types of visual perception with communicative meaning.

The dialog manager uses a short-term memory for 'active' information and a database that represents long-term memory, similar to Kawamura et al. [34]. In addition, we maintain a standard discourse model to represent communicative information. The short term memory is organized in chunks that contain semantic information, which is, like all other semantic data structures in our system, represented by typed feature structures (TFS). One part of the short-term memory is a user (focus) model that represents information about currently available communication partner(s). The user model collects information from different modalities about the user. The collected information is compared against database entries where the information either matches a single person, a set of persons, or allows creating a new entry.

The dialog manager applies a method that we call behavior selection. Each behavior defines a specific operation mode, and defines how the dialog manager interprets incoming events, how information is interpreted in the discourse, and which actions are taken by the strategy. In the given task we have found the following four states: idle, obtain attention, initiate conversation, conduct dialog. The full state-transition model is shown in figure 4. The state model contains a number of 'hard-coded' transitions. These are the transitions to idle if the system recognizes that the user leaves the conversation or after a timeout where no communication happens. The other transitions are defined from any state (i.e. idle, obtain attention and initiate conversation) to 'dialog', if the user has picked up conversation with the system. The remaining transitions are defined by the classifier and selection mechanism described in section 2. The behavior model also provides a basis for further extending the transitions influenced by motivations and drives.

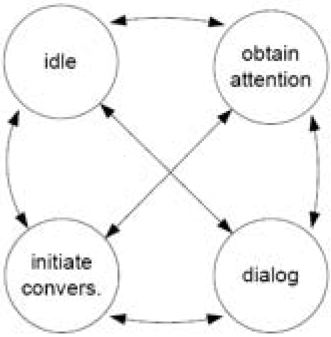

Fig. 4. State model with full transitions

4 Conclusions and Future Work

4.1 Conclusions

We have presented a first approach for a robot that learns to known persons completely on its own. The main challenges addressed are obtaining attention from the person, initiating a dialog, understanding new names input, building a face-ID database and recognizing persons, integration and decision-taking in dialog. The first experiments conducted show that such a system is possible with current technology, and motivate further development.

We have presented experiments on proactive behavior to obtain the user's attention and to initiate a dialog, so the robot doesn't need to wait for a person to initiate the dialog. During these experiments the robot has learned a behavior to obtain the attention and interest of persons walking by, without molesting other people, and a successive series of actions to initiate a dialog with interested persons.

We have further presented our speech recognizer with OOV recognition capabilities, as well as phoneme recognition and word spelling, and the integration of the face-ID recognizer as the main component for recognizing known and unknown persons prior to initiating a dialog. The dialog manager combines and coordinates the different components within the system. The different components have been evaluated individually, future experiments will show how the system operates in a real-world environment.

4.2 Future Works

In the future we plan to conduct further experiments with the fully integrated system, and evaluate the system in different environments and on different robots (especially our SFB588-robot Armar III) in out-of-the-lab scenarios. Further improvement is necessary for robust processing of the persons IDs - both vision and speech recognition can produce errors, so the database might contain different IDs for same person, or a mixture of different persons with the same ID. During the previous experiments, the system did not distinguish which persons to talk to. Selective dialogs will help in the future to remove this fortuity and to solve these errors by explicitly talking to specific persons. Furthermore, robustness issues and more elaborate methods to obtain better spelling - letters or phonemes - will be considered. The approach further motivates to obtain other personal information from the communication partners that exceed the name-only task.

Acknowledgments

This work was supported in part by the German Research Foundation (DFG) as part of the Collaborative Research Center 588 ”Humanoid Robots - Learning and Cooperating Multimodal Robots” and by the interACT research scholarship.

References

1. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. Robotics and Autonomous Systems 42 (2003) 143–166
2. Breazeal, C.: Social Interactions in HRI: The Robot View. IEEE Transactions on Man Cybernetics and Systems Vol. 34, Issue 2, (2004) 181–186

3. Byers, Z., Dixon, M., Smart, W.D., Grimm, C.M.: Say Cheese!: Experiences with a Robot Photographer. AI Magazine 25, Nr. 3 (2004) 37–46
4. Schulz, D., Burgard, W., Fox, D., Cremers, A.B.: Tracking Multiple Moving Targets with a Mobile Robot using Particle Filters and Statistical Data Association. Int. Conf. on Robotics and Automation (ICRA), IEEE Press, New Orleans (2001)
5. Lang, S., Kleinehagenbrock, M., Hohenner, S., Fritsch, J., Fink, G.A., Sagerer, G.: Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot. Proceedings of the Int. Conf. on Multimodal Interfaces, Vancouver, Canada, (2003) 28–35
6. Sidner, C., Kidd, C., Lee, C., Lesh, N.: Where to look: A study of human-robot engagement. Proceedings of the International Conference on Intelligent User Interfaces (IUI), ACM, (2004) 78–84
7. Dusan, S., Flanagan, J.: Adaptive Dialog Based upon Multimodal Language Acquisition. Proceedings of the Fourth Int. Conf. on Multimodal Interfaces, Pittsburgh, PA, USA, (2002)
8. Dusan, S., Flanagan J.: A System for Multimodal Dialogue and Language Acquisition. Invited, The 2nd Romanian Academy Conference on Speech Technology and Human-Computer Dialogue, Romanian Academy, Bucharest, Romania (2003)
9. Gorniak, P., Roy, D.: Probabilistic Grounding of Situated Speech using Plan Recognition and Reference Resolution. Proceedings of the Seventh International Conference on Multimodal Interfaces (2005)
10. Chung, G., Seneff, S.: Integrating Speech with Keypad Input for Automatic Entry of Spelling and Pronunciation of New Words. Proceedings of ICSLP'02, Denver, CO, USA (2002) 2061–2064
11. Chung, G., Seneff, S., Wang, C., Hetherington I.: A Dynamic Vocabulary Spoken Dialogue Interface. Proceedings of ICSLP'04, Jeju Island, Korea (2004)
12. Scharenborg, O., Seneff, S.: Two-pass strategy for handling OOVs in a large vocabulary recognition task. Proceedings of Interspeech'05 (2005) 1669–1672
13. Schaa, C.: Proaktive Initiierung von Dialogen für humanoide Roboter. Diploma Thesis, Universität Karlsruhe (2005)
14. Ekenel, H.K., Stiefelhagen, R.: Local Appearance based Face Recognition Using Discrete Cosine Transform. Proceedings of the 13th European Signal Processing Conference (EUSIPCO), Antalya, Turkey, (2005)
15. Ekenel, H.K., Stiefelhagen, R.: A Generic Face Representation Approach for Local Appearance based Face Verification. Proceedings of the CVPR IEEE Workshop on FRGC Experiments, San Diego, CA, USA (2005)
16. Ekenel, H.K., Stiefelhagen, R.: Analysis of Local Appearance-based Face Recognition on FRGC 2.0 Database. Face Recognition Grand Challenge Workshop (FRGC), Arlington, VA, USA (2006)
17. Ekenel, H.K., Pnevmatikakis, A.: Video-Based Face Recognition Evaluation in the CHIL Project - Run 1. Proceedings of the 7th Intl. Conf. on Automatic Face and Gesture Recognition (FG 2006), Southampton, UK (2006)
18. Ekenel, H.K., Jin, Q.: ISL Person Identification System in the CLEAR Evaluations. Proceedings of the CLEAR Evaluation Workshop, Southampton, UK (2006)
19. R. Snelick, M. Indovina: Large Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems. IEEE Trans. Pattern Anal. Mach. Intell., 27(3) (2005) 450–455
20. J. Kittler, M. Hatef, R. Duin, J. Matas: On Combining Classifiers. IEEE Trans. Pattern Anal. Mach. Intell., 20(3) (1998)
21. H. Holzapfel: Building multilingual spoken dialogue systems. Special issue of Archives of Control Sciences, G.eds. Z. Vetulani, 4 (2005)

22. H. Holzapfel and P. Gieselmann: A Way Out of Dead End Situations in Dialogue Systems for Human-Robot Interaction. Humanoids (2004)
23. M. Denecke: Rapid Prototyping for Spoken Dialogue Systems. Proc. of the 19th Int. Conf. on Computational Linguistics (COLING), Taiwan (2002)
24. Finke, M., P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal: The karlsruhe-verbmobil speech recognition engine. Proc. of ICASSP'97, Germany (1997)
25. Soltau, H., F. Metze, C. Fuegen, and A. Waibel: A one pass- decoder based on polymorphic linguistic context assignment. Proceedings of ASRU'01, Madonna di Campiglio, Trento, Italy (2001)
26. Fügen, C., Holzapfel, H., Waibel, A.: Tight coupling of speech recognition and dialog management - dialog-context grammar weighting for speech recognition. Proceedings of ICSLP'04, Jeju Island, Korea (2004)
27. Gieselmann, P., Holzapfel, H.: Multimodal Context Management within Intelligent Rooms. Proceedings of the 10th International Conference on Speech and Computer (SPECOM'05), Patras, Greece (2005)
28. Schaaf, T.: Erkennen und Lernen neuer Wörter. PhD Thesis, Universität Karlsruhe (2004)
29. Schaaf, T.: Detection Of OOV Words Using Generalized Word Models And A Semantic Class Language Model. Proceedings of Eurospeech (2001)
30. Hetherington, L.: A Characterization of the problem of new, out-of-vocabulary words in continuous speech recognition and understanding. Ph.D.-Thesis, MIT (1995)
31. Asadi, A., Schwartz, R., Makhoul, J.: Automatic detection of new words in a large-vocabulary continuous speech recognition system. Proceedings of ICASSP'90, IEEE Signal Processing Society, Albuquerque, New Mexico, USA (1990)
32. Hild, H.: Buchstabiererkennung mit neuronalen Netzen in Auskunftssystemen. Ph.D.-Thesis, Universität Karlsruhe, Fakultät für Informatik, Shaker Verlag, Aachen, ISBN 3-8265-3155-8, Karlsruhe, Germany (1997)
33. Holzapfel, H., Nickel, K., Stiefelhagen, R.: Implementation and Evaluation of a Constraint-Based Multimodal Fusion System for Speech and 3D Pointing Gestures. Proc. of the Int. Conf. on Multimodal Interfaces, State College, PA, USA (2004)
34. Kawamura, K.: Cognitive Approach to a Human Adaptive Robot Development. Proceedings of the Int. Workshop on Robot and Human Interactive Communication (RO-MAN), Nashville, TN, USA (2005)