Detection of Communication Partners from a Mobile Robot

S. Lang, M. Kleinehagenbrock, J. Fritsch, G. A. Fink, and G. Sagerer

Bielefeld University, Faculty of Technology, 33594 Bielefeld, Germany
slang@techfak.uni-bielefeld.de

Abstract. An important prerequisite for the natural interaction of humans with a mobile robot is the robot's capability to detect potential communication partners. In this paper we present an approach which uses a combination of person recognition and tracking with sound source localization realized in a multi-modal anchoring framework. As in open environments several potential communication partners can be present simultaneously we developed a rule-based method for selecting one specific person as the current communication partner.

1 Introduction

A prerequisite for the widespread use of mobile service robots in home and office environments is the development of systems with natural human-robot-interaction. While much research focuses on the communication process itself, it is also necessary to explore how robots can automatically recognize when and how long a user's attention is directed towards the robot for communication.

For this purpose some fundamental abilities of the robot are required. It must be able to detect persons in its vicinity and to track their movements over time. Additionally, as speech is the most important means of communication for humans, the detection and localization of sound sources is of great importance.

This paper is organized as follows: At first we discuss approaches that are related to the detection of communication partners in section 2. Then, in section 3 multi-modal anchoring is described. This is the basis of our approach for the detection of communication partners explained in section 4. The paper concludes with a short summary.

2 Related Work

As long as artificial systems interact with humans in static setups the detection of communication partners (CPs) can be achieved rather easily. For the interaction with an information kiosk the potential user has to enter a well defined space in front of the device (cf. e.g. [1]). In intelligent rooms usually the configuration of the sensors allows to monitor all persons involved in a meeting simultaneously (cf. e.g. [2]).

* This work has been supported by the German Research Foundation within the Collaborative Research Center 'Situated Artificial Communicators' and the Graduate Programs 'Task Oriented Communication' and 'Strategies and Optimization of Behavior'.

In Proc. 4th Workshop on Dynamic Perception, pages 183-188, Bochum, Germany, November 2002.
Akademische Verlagsgesellschaft Aka GmbH, Berlin.

In contrast to these scenarios a mobile robot does not act in a closed or even controlled environment. A prototypical application of such a system is its use as a tour guide in scientific laboratories or museums (cf. e.g. [3]). All humans approaching or passing the robot have to be considered to be potential CPs. In order to circumvent the problem of detecting humans in an unstructured and potentially changing environment in [3] a button on the robot itself has to be pushed to start the interaction.

The humanoid robots *SIG* [4] and *ROBITA* [5] currently demonstrate their capabilities in research labs. Both use a combination of visual face recognition and sound source localization for the detection of potential CPs. *SIG*'s focus of attention is directed towards the person currently speaking that is either approaching the robot or standing close to it. In addition to the detection of talking people *ROBITA* is also able to determine the addressee of spoken utterances. Thus it can distinguish speech directed towards itself from utterances spoken to another person. Both robots, *SIG* and *ROBITA*, can give feedback which person is currently considered to be the CP. *SIG* always turns its complete body towards the CP. *ROBITA* can use several combinations of body orientation, head orientation, and eye gaze to express different states of communication.

3 Anchoring

Person tracking with a mobile robot is a highly dynamic task. Due to motions of the tracked persons and of the robot itself the sensory perception of the persons is constantly changing. In order to control the robots behavior, connections between processes that work on the level of abstract representations of objects in the world (symbolic level) and processes that are responsible for the physical observation of these objects (sensory level) need to be established. These connections, called *anchors*, must be dynamic, since the same symbol must be connected to new percepts every time a new observation of the corresponding object is acquired.

We follow the definition of anchoring proposed in [6]: At every time step t, the anchor contains three elements: a symbol, which is used to denote an object, a percept of the same object, generated in the perceptual system, and a signature, meant to provide the estimate of the values of the observable properties of the object. If the anchor is grounded at time t, it contains the percept perceived at t as well as the updated signature. If the object is not observable at t and therefore the anchor is ungrounded, then no percept is stored in the anchor but the signature still contains the best available estimate.

3.1 Multi-Modal Anchoring

Anchoring as defined in [6] only considers the special case of connecting one symbol to the percepts acquired from one sensor. However, complex objects cannot be captured completely by a single sensor system alone. If more than one sensor is used, the symbolic description of an object has to be linked to different types of percepts, originating from different perceptual systems.

For this purpose we propose an approach for *multi-modal anchoring* [7]. It allows distributed anchoring of individual percepts from multiple modalities and copes with different spatio-temporal properties of the individual percepts. Every part of the complex object which is captured by one sensor is anchored by a single *component anchoring process*. The composition of all component anchors is realized by a *composite anchoring process* which establishes the connection between the symbolic description of the complex object and the percepts from the individual sensors. In addition to standard anchoring, the composite anchoring module requires a *composition model*, a *motion model*, and a *fusion model*:

- The composition model defines the spatial relationships of the components with respect to the composite object. It is used in the component anchoring processes to anchor only those percepts that satisfy the composition model.
- The motion model describes the type of motion of the complex object, and therefore allows to predict its position. Using the spatial relationships of the composition model, the position of percepts can be predicted, too. This information is used by the component anchoring processes in two ways: 1. If multiple percepts were generated from one perceptual system the component anchoring process selects the percept which is closest to the predicted position. 2. If the corresponding perceptual system receives its data from a movable sensor with a limited field of view (e.g. pan-tilt camera), it turns the sensor into the direction of the predicted position.
- The fusion model defines how the perceptual data from the component anchors has to be combined. It is important to note, that the processing time of the different perceptual systems may differ significantly. In this case the perceptual data is not received by the composition anchoring process in chronological order. For this purpose the composite anchor provides a chronologically sorted list of the fused perceptual data. New data from the component anchors is inserted in the list, and all subsequent entries are updated.

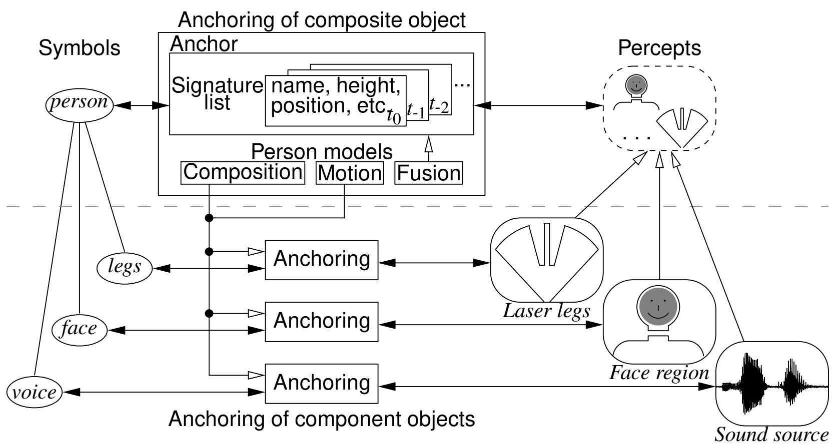

Fig. 1. Multi-modal anchoring of persons.

The detection of CPs from a mobile robot requires to track all persons in the vicinity of the robot. For this purpose we apply multi-modal anchoring. Our mobile robot is equipped with a laser range finder, a pan-tilt camera, and stereo microphones. Every sensor forms the basis for one perceptual system:

- The laser range finder covers a 180° field of view at a height of approximately 30cm. In range readings human pairs of legs result in a characteristic pattern that can be easily detected [7]. From a *legs*-percept distance and angle of the person relative to the robot are extracted.
- The camera is mounted on top of the robot at a height of 140cm. We developed a face detection method which copes with changing lighting conditions [8]. From a *face*-percept the distance, angle, height and identity of the person are extracted.
- Stereo microphones are applied to locate speakers using a method based on cross-powerspectrum phase analysis [9]. From a *voice*-percept the angle relative to the robot can be extracted.

The anchoring of a person is illustrated in Fig. 1. It is based on anchoring the three components *legs*, *face*, and *voice*.

3.2 Anchoring Multiple Persons

For the detection of CPs from a mobile robot usually more than one person has to be tracked at the same time. Then, several anchoring processes have to be run in parallel. In this case, multi-modal anchoring as described in the previous section may lead to the following conflicts between the individual composite anchoring processes:

1. A percept is selected by more than one anchoring process.
2. The anchoring processes try to control the pan-tilt unit of the camera in a contradictory way.

To resolve these problems a *supervising module* is required, which controls the selection of percepts and the access to the pan-tilt camera.

To handle the first problem, the supervising module restricts the access to the pan-tilt unit of the camera to only one composite anchoring process at a time. How access is granted to the processes depends on the intended application. An example is given for the detection of CPs in the following section.

In order to avoid the second problem, the selection of percepts is implemented as follows. Instead of selecting a specific percept deterministically every component anchoring process assigns scores to all percepts rating the proximity to the predicted position. Subsequently, the supervising module computes the optimal non-contradictory assignment of percepts to component anchors. Percepts that are not assigned to any of the existing anchoring processes are used to establish new anchors. Additionally, an anchor that was not updated for a certain period of time will be removed by the supervising module.

4 Detection of Communication Partners

For the detection of CPs from a mobile system we apply multi-modal anchoring of persons, as described in the previous sections. Every person in the vicinity of the robot is anchored by one anchoring process. From the anchoring processes the following attributes can be extracted: *standing*: The last positions of a person are known; it can therefore be decided, whether a person is walking or standing still. *speaking*: From the microphones it is known, whether a person is speaking or is silent. *facing*: The face anchoring process provides information, whether a person is facing the robot or looking in a different direction.

Note, that due to the limited field of view of the camera the attribute *facing* can not be computed for all persons simultaneously. The control of the access to the pan-tilt unit of the camera by the supervising module has an important relevance for this application.

For CP detection we propose the following set of heuristic rules, that are based on the three above mentioned attributes *standing*, *speaking*, and *facing*:

- A person that is not standing but walking is considered as a passer-by, and is therefore definitely no CP (rule 1).
- Whether a person standing still that is also speaking is classified as CP depends on the orientation of the head:
  - A person facing the robot is classified as CP (rule 2).
  - A person not facing the robot is assumed to be talking to someone else than the robot (e.g. another person) and therefore is definitely no CP (rule 3).
  - If no information from the camera is available, no classification is possible, so the person is a potential CP (rule 4).

The remaining three configurations of attributes (*standing*, not *speaking*, and any state of *facing*) leave the person's state unchanged (rule 5). This means that a person which was previously recognized as CP will be still considered as CP.

The rules for the detection of CPs are now used to define the behavior of the robot. On the one hand, the robot should direct its attention towards the person which was recognized as CP. This is done by turning the front of the robot into the direction of the CP, standing *face-to-face*. The anchoring process corresponding to that person gets access to the pan-tilt camera and keeps the person in the center of the field of view. On the other hand, the robot must be able to recognize a new CP, when the current CP is not speaking (rule 5). Only a person that is speaking can take over the role of the CP. If a person speaking is in the field of view of the camera one of the rules 2 or 3 can be applied and a decision is possible. If a person speaking is not in the field of the camera it is considered as a potential CP (rule 4). Then, the corresponding anchoring process gets access to the pan-tilt camera in order to focus the potential CP. Now a decision can be made according to rules 2 and 3. If the person is facing the robot it becomes the new CP, otherwise the anchoring process of the old CP again gets back access to the pan-tilt camera. Note, that while the camera is used to check the state of other persons the front of the robot is still directed towards the current CP, thus signaling that this person is the current CP. A sample behavior of the robot is depicted in Fig. 2.

5 Summary

We presented an approach for the detection of communication partners (CPs) from a mobile robot. The detection requires to simultaneously track persons in the vicinity of

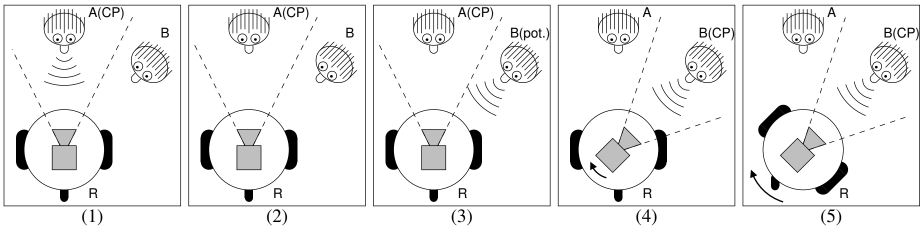

Fig. 2. Sample behavior with two persons standing near the robot R. In (1) person A is the CP, thus the robot directs its attention towards A. Then A stops speaking but remains the CP (2). In (3) person B begins to speak. Unless B's head is not in the camera's field of view, B is a potential CP. Therefore the robot turns the camera into the direction of B, still showing A its attention (4). Since person B is facing the robot, B becomes the new CP, and the robot turns towards B in (5).

the robot. This is achieved by multi-modal anchoring based on three types of sensors: pan-tilt camera, laser range finder, and stereo microphones. The anchoring processes provide three person attributes: standing, speaking, facing. We developed a set of heuristic rules which define if a person is considered as a CP. In addition, the competing access of the anchoring processes to the pan-tilt unit of the camera is described.

References

1. V. Pavlović, A. Garg, J. Rehg, and T. Huang. Multimodal speaker detection using error feedback dynamic bayesian networks. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR-00)*, pages 34–43, Los Alamitos, June 2000.
2. R. Stiefelhagen, J. Yang, and A. Waibel. Estimating focus of attention based on gaze and sound. In *Workshop on Perceptive User Interfaces (PUI'01)*, November 2001.
3. W. Burgard, A. B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. The interactive museum tour-guide robot. In *Proc. of the Fifteenth National Conf. on Artificial Intelligence (AAAI-98)*, 1998.
4. H. G. Okuno, K. Nakadai, and H. Kitano. Social interaction of humanoid robot based on audio-visual tracking. In *Proc. of 18th Intern. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-2002)*, June 2002.
5. Y. Matsusaka, S. Fujie, and T. Kobayashi. Modeling of conversational strategy for the robot participating in the group conversation. In *Proc. European Conf. on Speech Communication and Technology*, pages 2173–2176, Aalborg, Denmark, September 2001.
6. S. Coradeschi and A. Saffiotti. Perceptual anchoring of symbols for action. In *Proc. of the 17th IJCAI Conf.*, pages 407–412, Seattle, WA, 2001.
7. M. Kleinehagenbrock, S. Lang, J. Fritsch, F. Lömker, G. A. Fink, and G. Sagerer. Person tracking with a mobile robot based on multi-modal anchoring. In *IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN)*, 2002. to appear.
8. J. Fritsch, S. Lang, M. Kleinehagenbrock, G. A. Fink, and G. Sagerer. Improving adaptive skin color segmentation by incorporating results from face detection. In *IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN)*, 2002. to appear.
9. D. Giuliani, M. Omologo, and P. Svaizer. Talker localization and speech recognition using a microphone array and a cross-powerspectrum phase analysis. In *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 1243–1246, Yokohama, Japan, 1994.