Providing the Basis for Human-Robot-Interaction: A Multi-Modal Attention System for a Mobile Robot

Sebastian Lang, Marcus Kleinehagenbrock, Sascha Hohenner,
Jannik Fritsch, Gernot A. Fink, and Gerhard Sagerer
Bielefeld University, Faculty of Technology, Bielefeld, Germany
{slang, mkleineh, sascha, jannik, gernot, sagerer}@techfak.uni-bielefeld.de

ABSTRACT

In order to enable the widespread use of robots in home and office environments, systems with natural interaction capabilities have to be developed. A prerequisite for natural interaction is the robot's ability to automatically recognize when and how long a person's attention is directed towards it for communication. As in open environments several persons can be present simultaneously, the detection of the communication partner is of particular importance. In this paper we present an attention system for a mobile robot which enables the robot to shift its attention to the person of interest and to maintain attention during interaction. Our approach is based on a method for multi-modal person tracking which uses a pan-tilt camera for face recognition, two microphones for sound source localization, and a laser range finder for leg detection. Shifting of attention is realized by turning the camera into the direction of the person which is currently speaking. From the orientation of the head it is decided whether the speaker addresses the robot. The performance of the proposed approach is demonstrated with an evaluation. In addition, qualitative results from the performance of the robot at the exhibition part of the ICVS'03 are provided.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Sensor fusion, Tracking*; H.1.2 [**Models and Principles**]: User/Machine Systems; I.5.5 [**Pattern Recognition**]: Implementation—*Interactive systems*

General Terms

Algorithms, Performance, Experimentation

Keywords

Human-robot-interaction, Multi-modal person tracking, Attention

1. INTRODUCTION

A prerequisite for the successful application of mobile service robots in home and office environments is the development of systems with natural human-robot-interfaces. Much research focuses on the communication process itself, e.g. speaker-independent speech recognition or robust dialog systems. In typical tests of such human-machine interfaces, the presence and position of the communication partner is known beforehand as the user either wears a close-talking microphone or stands at a designated position. On a mobile robot that operates in an environment where several people are moving around, it is not always obvious for the robot which of the surrounding persons wants to interact with it. Therefore, it is necessary to develop techniques that allow a mobile robot to automatically recognize when and how long a user's attention is directed towards it for communication.

For this purpose some fundamental abilities of the robot are required. First of all, it must be able to detect persons in its vicinity and to track their movements over time in order to differentiate between persons. In previous work, we have demonstrated how tracking of persons can be accomplished using a laser range finder and a pan-tilt color camera [6].

As speech is the most important means of communication for humans, we extended this framework to incorporate sound source information for multi-modal person tracking and attention control. This enables a mobile robot to detect and localize sound sources in the robot's surroundings and, therfore, to observe humans and to shift its attention to a person that is likely to communicate with the robot. The proposed attention system is part of a larger research effort aimed at building BIRON – the Bielefeld Robot Companion.

Figure 1: Even in crowded situations (here at the ICVS'03) the mobile robot BIRON is able to robustly track persons and shift its attention to the speaker.

ICMI'03, November 5–7, 2003, Vancouver, British Columbia, Canada.
Copyright 2003 ACM 1-58113-621-8/03/0011 ...$5.00.

BIRON has already performed attention control successfully during several demonstrations. Figure 1 depicts a typical situation during the exhibition of our mobile robot at the International Conference on Computer Vision Systems (ICVS) 2003 in Graz.

The 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 the robot hardware is presented. Next, multi-modal person tracking is outlined in section 4, followed by the explanation of the corresponding perceptual systems in section 5. This is the basis of our approach for the detection of communication partners explained in section 6. In section 7 an extensive evaluation of the system is presented. The paper concludes with a short summary in section 8.

2. RELATED WORK

As long as artificial systems interact with humans in static setups the detection of communication partners can be achieved rather easily. For the interaction with an information kiosk the potential user has to enter a definite space in front of it (cf. e.g. [14]). In intelligent rooms usually the configuration of the sensors allows to monitor all persons involved in a meeting simultaneously (cf. e.g. [18]).

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 communication partners. In order to circumvent the problem of detecting humans in an unstructured and potentially changing environment, in the approach presented in [3] a button on the robot itself has to be pushed to start the interaction.

Two examples for robots with advanced human-robot interfaces are *SIG* [13] and *ROBITA* [12] which currently demonstrate their capabilities in research labs. Both use a combination of visual face recognition and sound source localization for the detection of a person of interest. *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 communication partner. *SIG* always turns its complete body towards the person of interest. *ROBITA* can use several combinations of body orientation, head orientation, and eye gaze.

The multi-modal attention system for a mobile robot presented in this paper is based on face recognition, sound source localization and leg detection. In the following related work on these topics will be reviewed.

For human-robot interfaces tracking of the user's face is indispensable. It provides information about the user's identity, state, and intent. A first step for any face processing system is to detect the locations of faces in the robot's camera image. However, face detection is a challenging task due to variations in scale and position within the image. In addition, it must be robust to different lighting conditions and facial expressions. A wide variety of techniques has been proposed, for example neural networks [15], deformable templates [23], skin color detection [21], or principle component analysis (PCA), the so-called Eigenface method [19]. For an overview the interested reader is referred to [22, 9].

In current research on sound or speaker localization mostly microphone arrays with at least 3 microphones are used. Only a few approaches employ just one pair of microphones. Fast and robust techniques for sound (and therefore speaker) localization are e.g. the Generalized Cross-Correlation Method [11] or the Cross-Powerspectrum Phase Analysis [8], which both can be applied for microphone-arrays as well as for only one pair of microphones. More complex algorithms for speaker localization like spectral separation and measurement fusion [2] or Linear-Correction Least-Squares [10] are also very robust and can additionally estimate the distance and the height of a speaker or separate different audio sources. Such complex algorithms require more than one pair of microphones to work adequately and also require substantial processing power.

In mobile robotics 2D laser range finders are often used, primarily for robot localization and obstacle avoidance. A laser range finder can also be applied to detect persons. In the approach presented in [16] for every object detected in a laser scan features like diameter, shape, and distance are extracted. Then, fuzzy logic is used to determine which of the objects are pairs of legs. In [17] local minima in the range profile are considered to be pairs of legs. Since other objects (e.g. trash bins) produce patterns similar to persons, moving objects are distinguished from static objects, too.

3. ROBOT HARDWARE

The hardware platform for BIRON is a Pioneer PeopleBot from ActivMedia (Fig. 2) with an on-board PC (Pentium III, 850 MHz) for controlling the motors and the on-board sensors and for sound processing. An additional PC (Pentium III, 500 MHz) inside the robot is used for image processing.

The two PC's running Linux are linked with a 100 Mbit Ethernet and the controller PC is equipped with wireless Ethernet to enable remote control of the mobile robot. For the interaction with a user a 12" touch screen display is provided on the robot.

A pan-tilt color camera (Sony EVI-D31) is mounted on top of the robot at a height of 141 cm for acquiring images of the upper body part of humans interacting with the robot. Two AKG far-field microphones which are usually used for hands free telephony are located at the front of the upper platform at a height of 106 cm, right below the touch screen display. The distance between the microphones is 28.1 cm. A SICK laser range finder is mounted at the front at a height of approximately 30 cm.

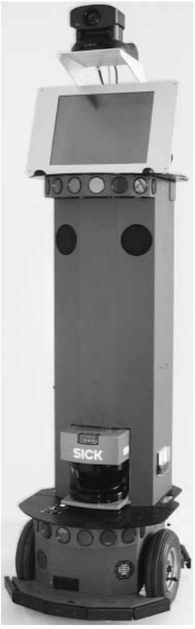

Figure 2: The mobile robot BIRON.

For robot navigation we use the ISR (Intelligent Service Robot) control software developed at the Center for Autonomous Systems, KTH, Stockholm [1].

4. MULTI-MODAL PERSON TRACKING

In order to enable a robot to direct its attention to a specific person it must be able to distinguish between different persons. Therefore, it is necessary for the robot to track all persons present as robustly as possible.

Person tracking with a mobile robot is a highly dynamic task. As both, the persons tracked and the robot itself might be moving the sensory perception of the persons is constantly changing. Another difficulty arises from the fact that a complex object like a person usually cannot be captured completely by a single sensor system alone. Therefore, we use the sensors presented in section 3 to obtain different percepts of a person:

- The camera is used to recognize faces. From a face detection step the distance, direction, and height of the observed person are extracted, while an identification step provides the identity of the person if it is known to the system beforehand (see section 5.1).

- Stereo microphones are applied to locate sound sources using a method based on Cross-Powerspectrum Phase Analysis [8]. From the result of the analysis the direction relative to the robot can be estimated (see section 5.2).

- The laser range finder is used to detect legs. In range readings pairs of legs of a human result in a characteristic pattern that can be easily detected [6]. From detected legs the distance and direction of the person relative to the robot can be extracted (see section 5.3).

The processes which are responsible for processing the data of these sensors provide information about the same overall object: the person. Consequently, this data has to be fused. We combine the information from the different sensors in a multi-modal framework which is described in the following section.

4.1 Multi-Modal Anchoring

In order to solve the problem of person tracking we apply *multi-modal anchoring* [6]. This approach extends the idea of standard anchoring as proposed in [4]. The goal of anchoring is defined as establishing 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). 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.

Therefore, in standard anchoring at every time step t, an anchor contains three elements: a symbol, which is used to denote an object, a percept of the same object, generated by the corresponding perceptual system, and a signature, which is meant to provide an estimate for 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.

Because standard anchoring only considers the special case of connecting one symbol to the percepts acquired from one sensor, the extension to multi-modal anchoring was necessary in order to handle data from several sensors. Multi-modal anchoring allows to link the symbolic description of a complex object to different types of percepts, originating from different perceptual systems. It enables 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 the domain of person tracking the person itself is the composite object while its components are face, speech, and legs, respectively. In addition to standard anchoring, the composite anchoring module requires a *composition model*, a *motion model*, and a *fusion model*:

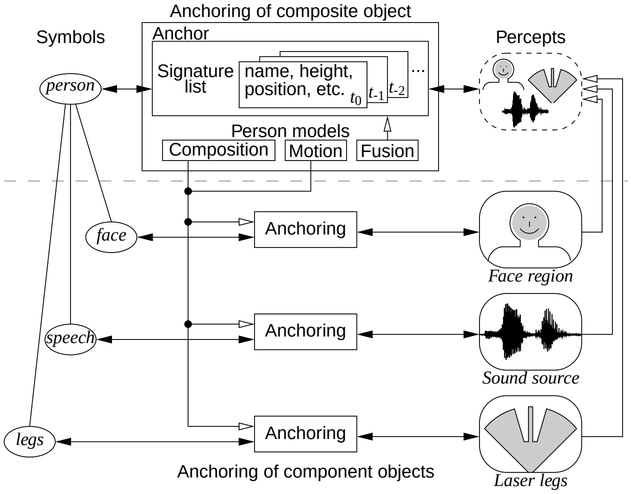

Figure 3: Multi-modal anchoring of persons.

- 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 are 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 steerable 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 times of the different perceptual systems may differ significantly. Therefore, the perceptual data may not arrive at the composite anchoring process in chronological order. Consequently, 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.

The anchoring of a single person is illustrated in Figure 3. It is based on anchoring the three components *legs*, *face*, and *speech*. For more details please refer to [6].

4.2 Tracking Multiple Persons

If more than one person has to be tracked simultaneously, 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:

- The anchoring processes try to control the pan-tilt unit of the camera in a contradictory way.

- A percept is selected by more than one anchoring process.

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

The first problem is handled in the following way: 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. For the task of detecting communication partners which is presented in this paper, only the anchoring process corresponding to the currently selected person of interest controls the pan-tilt unit of the camera (see section 6).

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. After all component anchoring processes have assigned scores, 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.

5. PERCEPTUAL SYSTEMS

In order to supply the anchoring framework presented in 4.1 with sensory information about observed persons, three different perceptual systems are used. These are outlined in the following subsections.

5.1 Face Recognition

In our previous work [6], face detection was realized using a method which combines adaptive skin-color segmentation with face detection based on Eigenfaces [7]. The segmentation process reduces the search space, so that only those sub-images which are located at skin colored regions have to be verified with the Eigenface method. In order to cope with varying lighting conditions the model for skin-color is continuously updated with pixels extracted from detected faces. This circular process requires initialization, which is realized by performing face detection using Eigenfaces on the whole image, since initially no suitable model for skin-color is available. This method has two major drawbacks: It is very sensitive to false positive detections of faces, since then the skin-model may adapt to a wrong color. In addition, initialization is computationally very expensive.

In our current system presented in this paper, the detection of faces (in frontal view) is based on the framework proposed by Viola and Jones [20]. This method allows to process images very rapidly with high detection rates for the task of face detection. Therefore, neither a time consuming initialization nor the restriction of the search using a model of skin color is necessary.

The detection is based on two types of features (Fig. 4), which are the same as proposed in [24]. A feature is a scalar value which is computed by the weighted sum of all intensities of pixels in rectangular regions. The computation can be realized very efficiently using integral images (see [20]). The features have six degrees of freedom for two-block features (x, y, w, h, dx, dy) and seven degrees of freedom for three-block features $(x, y, w, h, dx, dy, dx')$. With restrictions to the size of the rectangles and their distances we obtain about 300.000 different features for sub-windows of a size of 20 × 20 pixels. Classifiers are constructed by selecting a small number of important features using AdaBoost [5]. A cascade of classifiers $(C_1, \ldots, C_n)$ of increasing complexity (increasing number of features) forms the over-all face detector (Fig. 5). For face detection an image is scanned, and every sub-image is classified with the first classifier C_1 of the cascade. If classified as non-face, the process continues with the next sub-image. Otherwise the current sub-image is passed to the next classifier (C_2) and so on.

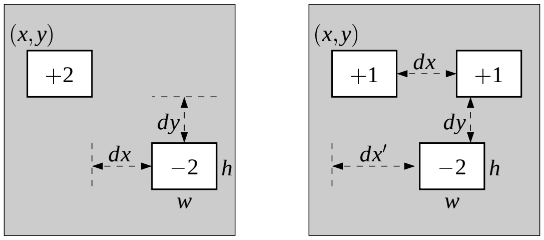

Figure 4: The two types of features used for face detection. Each feature takes a value which is the weighted sum of all pixels in the rectangles.

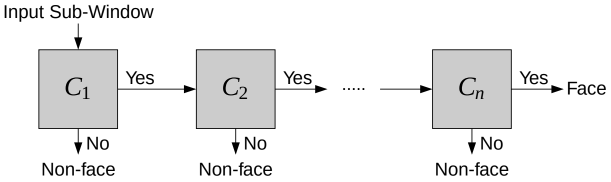

Figure 5: A cascade of n classifiers of increasing complexity enables fast face detection.

The first classifier of the cascade is based on only two features, but rejects approximately 75 % of all sub-images. Therefore, the detection process is very fast. The cascade used in our system consists of 16 classifiers based on 1327 features altogether.

In order to update the multi-modal anchoring process the position of the face is extracted: With the orientation of the pan-tilt camera, the angle of the face relative to the robot is calculated. The size of the detected face is used to estimate the distance of the person: Assuming that sizes of heads of adult humans only vary to a minor degree, the distance is proportional to the reciprocal of the size. The height of the face above the ground is also extracted by using the distance and the camera position.

Since the approach presented so far does not provide face identification, a post-processing step is is required. Therefore, we use a slightly enhanced version of the Eigenface method [19]. Each individual is represented in face space by a mixture of several Gaussians with diagonal covariances. Practical experiments have shown that the use of four to six Gaussians leads to a satisfying accuracy in discriminating between a small set of known persons.

5.2 Sound Source Localization

In order to detect speaking persons, we realize the localization of sound sources using a pair of microphones. Given a sound source s in 3D space, the distances d_1 and d_2 between s and the two microphones m_1 and m_2 generally differ by the amount of $\Delta d := d_2 - d_1$ (see Fig. 6). This difference Δd results in a time delay δ of the received signal between the left and the right channel (microphone). Based on Cross-Powerspectrum Phase Analysis [8] we first calculate a spectral correlation measure

$$C(\tau) = FT^{-1}\left(\frac{\hat{S}_L(f)\,\hat{S}^*_R(f)}{|\hat{S}_L(f)|\,|\hat{S}_R(f)|}\right) \tag{1}$$

where $\hat{S}_L(f)$ and $\hat{S}_R(f)$ are the short-term power spectra of the left and the right channel, respectively (calculated within a 43 ms window from the signal sampled at 48 kHz). If only a single sound

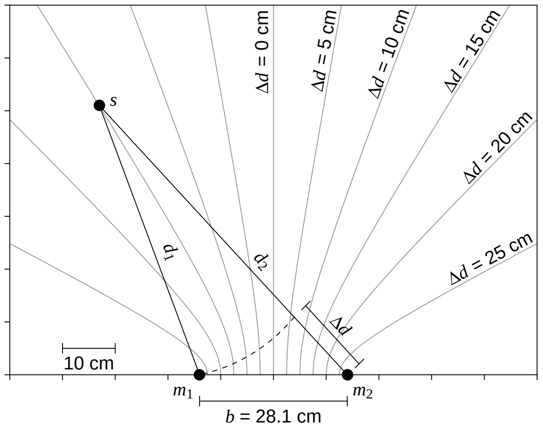

Figure 6: The distances d_1 and d_2 between the sound source s and the two microphones m_1 and m_2 differ by the amount of Δd. All sound events with identical Δd are located on one half of a two-sheeted hyperboloid (gray).

source is present the time delay δ will be given by the argument τ that maximizes the spectral correlations measure $C(\tau)$:

$$\delta = \arg\max_{\tau} \; C(\tau) \tag{2}$$

Taking into account also local maxima delivered by equation (1), we are able to detect several sound sources simultaneously.

Even in the planar case, where all sound sources are on the same level as the microphones, the position of s can be estimated only if its distance is known or additional assumptions are made. In a simplified geometry the microphone distance b is considered sufficiently small compared to the distance of the source. Therefore, the angles of incidence of the signals observed at the left or right microphone, respectively, will be approximately equal and can be calculated directly from Δd. In the 3D-case the observed time delay not only depends on the direction and distance but also on the relative elevation of the source with respect to the microphones. Therefore, given only Δd the problem is under-determined.

All sound events which result in the same Δd are located on one half of a two-sheeted hyperboloid, given by

$$\frac{s_x^2 + s_z^2}{\frac{1}{4}\left(b^2 - (\Delta d)^2\right)} - \frac{s_y^2}{\frac{1}{4}(\Delta d)^2} = -1 \tag{3}$$

where (s_x, s_y, s_z) is the position of the sound source given in Cartesian coordinates. The axis of symmetry of the hyperboloid coincides with the axis on which the microphones are located (y-axis). Figure 6 shows the intersections of hyperboloids for different Δd with the plane spanned by s, m_1, and m_2. Consequently, the localization of sound sources in 3D using two microphones requires additional information.

As in our scenario sound sources of interest correspond to persons talking, the additional spatial information necessary can be obtained from the other perceptual systems of the multi-modal anchoring framework. Leg detection and face recognition provide information about the direction, distance, and height of a person with respect to the local coordinate system of the robot. Even if no face was detected at all, the height of a person can be estimated as the standard size of an adult.

In order to decide whether a sound percept can be assigned to a specific person, the sound source has to be located in 3D. For this purpose it is assumed that the sound percept originates from the person and is therefore located at the same height and same distance. Then, the corresponding direction of the sound source can be calculated from equation (3) transformed to cylindric coordinates. Depending on the difference between this direction and the direction in which the person is located, the sound percept is assigned to the person's sound anchor. Similar to other component anchors, the direction of the speech is also fused with the position of the person. Note that the necessity of positional attributes of a person for the localization of speakers implies that speech can not be anchored until the legs or the face of a person have been anchored.

In conclusion, the use of only one pair of microphones is sufficient for feasible speaker localization in the multi-modal anchoring framework.

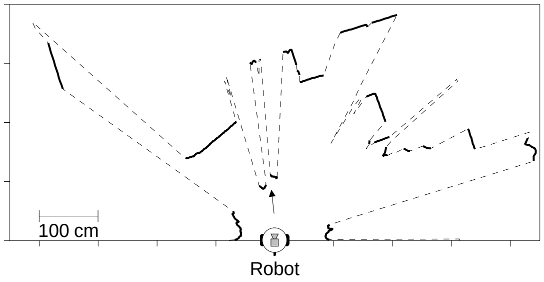

Figure 7: A sample laser scan. The arrow marks a pair of legs.

5.3 Leg Detection

The laser range finder provides distance measures within a 180° field of view at leg-height. The angular resolution is 0.5° resulting in 361 reading points for a single scan (see Fig. 7 for an example). Usually, human legs result in a characteristic pattern which can be easily detected. This is done as follows: At first, neighboring reading points with similar distance values are grouped into segments. Then, these segments are classified as legs or non-legs based on a set of features (see [6]). Finally, legs with a distance that is below a threshold are grouped into pairs of legs.

6. FOCUSING THE ATTENTION

For the detection of a person of interest from our mobile robot we apply multi-modal person tracking, as described in section 4. Every person in the vicinity of the robot is anchored and, therefore, tracked by an individual anchoring process, as soon as the legs or the face can be recognized by the system.

If the robot detects that a person is talking, this individual becomes the person of interest and the robot directs its attention towards it. This is achieved by turning the camera into the direction of the person. The anchoring process corresponding to the person of interest maintains access to the pan-tilt camera and keeps the person in the center of the camera's field of view. If necessary, the entire robot basis is turned in the direction of the person of interest. If this person moves to far away from the robot, the robot will start to follow the person. This behavior ensures that the sensors of the robot do not loose track of this person. Moreover, the person can guide the robot to a specific place.

As long as the speech of the person of interest is anchored, other people talking are ignored. This allows the person of interest to take breath or make short breaks while speaking without loosing the robots attention. When the person of interest stops talking for more than two seconds, the person of interest looses its *speech* anchor. Now, another person can become the person of interest. If no other person is speaking in the vicinity of the robot, the person which

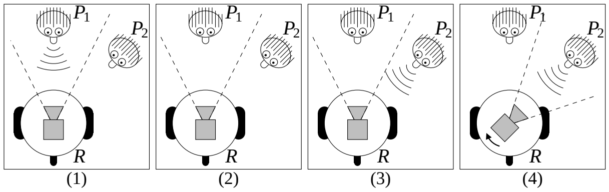

Figure 8: Sample behavior with two persons P_1 and P_2 standing near the robot R: In (1) P_1 is considered as communication partner, thus the robot directs its attention towards P_1. Then P_1 stops speaking but remains person of interest (2). In (3) P_2 begins to speak. Therefore the robot's attention shifts to P_2 by turning its camera (4). Since P_2 is facing the robot, P_2 is considered as new communication partner.

is in the focus of attention of the robot remains person of interest. Only a person that is speaking can take over the role of the person of interest. Notice, that a person which is moving fast in front of the robot is considered as a passer-by, and hence is definitely no person of interest even if this person is speaking.

In addition to the attention system described so far, which enables the robot to detect the person of interest and to maintain its attention during interaction, the robot decides whether the person of interest is addressing the robot and, therefore, is considered as communication partner. This decision is based on the orientation of the person's head, as it is assumed that humans face their addressees for most of the time while they are talking to them. Whether a tracked person faces the robot or not is derived from the face recognition system. If the face of the person of interest is detected for more than 20 % of the time the person is speaking, this person is considered to be the communication partner.

A sample behavior of the robot is depicted in Figure 8.

7. SYSTEM PERFORMANCE

In order to analyze the performance of the proposed approach, we present results from three different types of evaluation. At first, we study the accuracy of sound source localization independently. The second part deals with a quantitative evaluation of our approach for a multi-modal attention system. Finally, qualitative results from a performance of the robot at the exhibition part of the ICVS'03 are presented.

7.1 Evaluation of Sound Source Localization

The objective of this evaluation was to analyze the accuracy of locating speakers with a pair of microphones using the method described in section 5.2 independently from the multi-modal anchoring framework. In order to be able to estimate the arrival angle relative to the microphones, the setup for the experiment was arranged such that the sound source (mouth of the speaker) was always at the same height as the microphones. Therefore, the simplified geometric model mentioned in section 5.2 can be used.

The experiments were carried out with five subjects. Every subject was positioned at six different angles (0°, 10°, 20°, 40°, 60°, and 80°), and at two different distances (100 cm and 200 cm), respectively, resulting in 12 positions altogether. At every position a subject had to read out one specific sentence which took about 8 seconds. During every utterance the position of the speaker was calculated every 50 ms.

Based on the angles estimated by our localization algorithm we calculated the mean angle and the variance for every speaker. It is important to note, that in our setup it is almost impossible to position the test speaker accurately on the target angle. For this reason, we used the mean estimated angle for every speaker instead of the target angle to calculate the variance. Following the calculation of mean angle and variance for every speaker we averaged for every position the mean angle and the variance across all speakers.

Angle	Distance between speaker and robot			
	100 cm		200 cm	
	μ	σ	μ	σ
0°	-0.9°	0.56	-0.3°	0.81
10°	9.1°	0.34	9.2°	0.37
20°	18.9°	0.21	19.3°	0.27
40°	38.2°	0.50	38.8°	0.22
60°	57.7°	0.40	57.5°	0.64
80°	74.0°	2.62	73.3°	2.18

Table 1: Averaged estimated speaker positions μ and averaged variances σ for the acoustic speaker localization.

Table 1 shows the results of our experiments. First, the results suggest that the robot was not correctly aligned, because especially for small angles (0° to 20°) the averaged angle differs constantly from the target angle about 1°. Under this justifiable assumption the speaker localization works very well for angles between 0° and 40°. The estimated angle is nearly equivalent to the actual angle and the variance is also very low. Furthermore, the acoustic position estimation works equally well for 100 cm and for 200 cm. For angles greater than 40° the difference between estimated angle and target angle as well as the variance increases. This means that the accuracy of the acoustic position estimation decreases with an increasing target angle. The main reason for this behavior is the directional characteristic of the microphones.

However, the evaluation has shown that the time delay estimation works reasonably well. Thus the sound source localization provides important information for the detection and localization of the current person of interest.

7.2 Evaluation of the Attention System

The objective of this evaluation was to analyze the performance of the attention system presented in this paper. On the one hand, the capability of the robot to successfully shift its attention on the speaker, and to recognize when it was addressed was investigated. On the other hand, details about the perceptual sub-systems were of interest.

The experiment was carried out in an office room (Fig. 9). Four persons were standing around the robot at designated positions. In reference to the local coordinate system of the robot, person P_1 was located at a distance of 120 cm and an angle of 45°, where 0° is defined as the direction ahead of the robot. Person P_2 was located at $(140 \text{ cm}, 0^\circ)$, person P_3 at $(180 \text{ cm}, -30^\circ)$, and person P_4 at $(160 \text{ cm}, -60^\circ)$. The subjects were asked to speak for about 10 seconds, one after another. They had to either address the robot or one of the other persons by turning their heads into the corresponding direction. There were no restrictions on how to stand. The order in which the persons were speaking was predetermined (see Table 2). The experiment was carried out three times with nine different subjects altogether.

The following results were achieved:

- The attention system was always able to determine the correct person of interest within the time the person was speaking. However, in some situations either the reference to the

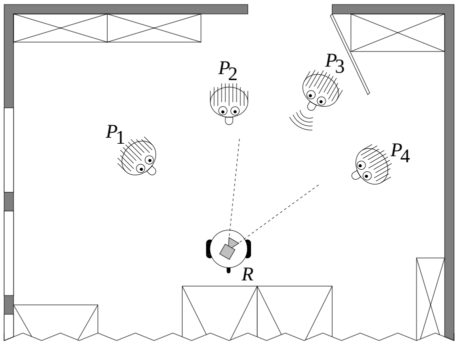

Figure 9: Setup for the evaluation of the attention system.

Step Person	1	2	3	4	5	6	7	8	9	10	11	12
P_1	🤖					👤			🤖			
P_2		🤖					👤					🤖
P_3			🤖		👤						🤖	
P_4				🤖				👤		🤖		

Table 2: Order in which the persons were speaking, either to the robot (steps 1–4 and 9–12) or to another person (steps 5–8).

last person of interest was sustained too long or an incorrect person of interest was selected intermediately. A diagram of the robot's focus of attention is shown in Figure 10. The erroneous behavior occurred in 4 of the 36 time slices: In these cases, the robot shifted its attention to a person which was currently not speaking (see column 5 in all experiments and column 4 in the last experiment in Fig. 10). Note that in all failure cases person P_2, which was located in front of the robot, was selected as person of interest. In addition, there were two shifts which were correct but had a very long delay (eighth time slice of the first and the third experiment). Again, the person in front of the robot (P_2) was involved. All errors occurred because a sound source was located in the direction of 0°, although person P_2 was not speaking. This can be explained with the noise of the robot itself, which is interpreted as a sound source in the corresponding direction. This error could be suppressed using voice activity detection, which distinguishes speech from noise. This will be part of our future work.

As the diagram in Fig. 10 shows, every shift of attention had a delay of approximately 2 seconds. This results from the anchoring framework: The anchor for the sound source is removed after a period of 2 seconds with no new assigned percepts. Now, if another person is talking it becomes the person of interest.

- The decision whether the current person of interest was addressing the robot or not was made as described in section 6. It was correct for all persons in all runs. This means that the robot always determined himself as addressee in steps 1–4 and 9–12, and never in steps 5–8.

These results prove that the presented approach for a multi-modal attention system on a mobile robot is capable to identify communication partners successfully.

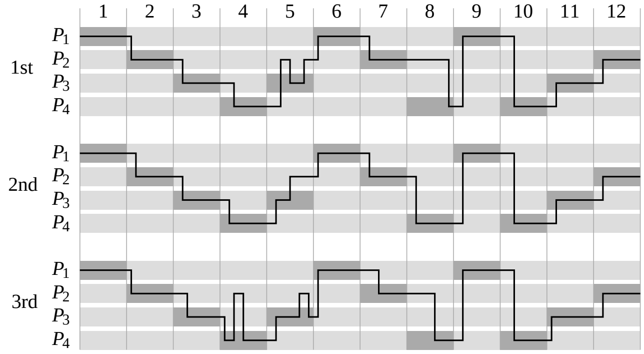

Figure 10: Diagram for the three runs of the experiment. Every person is assigned a track (light-gray) which is shaded while the person was speaking. The solid line shows which person was in focus of the robot's attention.

In addition the following measurements concerning the anchoring framework were extracted during the experiments: The attention system and the face recognition were running on one PC (Pentium III, 500 MHz), while the sound source localization and the robot control software were running on the other PC (Pentium III, 850 MHz). Face recognition was performed on images of a size of 256×192 at a rate of 9.6 Hz. Localization of sound sources was running at a rate of 5.5 Hz. The laser range finder provided new data at a rate of 4.7 Hz while the processing time for the detection of legs was negligible.

The anchoring processes of the persons which were currently speaking to the robot were updated with percepts at a rate of 15.4 Hz. Face percepts were assigned to the corresponding anchor at 71.4 % of the time. Note, that after a new person of interest is selected it takes up to approximately 1 second until the camera is turned and the person is in the field of view. During this time, no face percept for the person of interest can be generated. Sound percepts were assigned at 69.5 % of the time, and leg percepts at 99.9 % of the time.

The multi-modal anchoring framework was able to quantify the body heights of all subjects with an accuracy of at least ±5 cm, which was sufficient to precisely locate sound sources in 3D (see section 5.2).

7.3 Performance at an Exhibition

In the beginning of April 2003 our robot was presented at the exhibition part of the International Conference on Computer Vision Systems (ICVS) in Graz. There we were able to demonstrate the robot's capabilities in multi-modal person tracking, and also in following people. BIRON was continuously running without any problems.

On the two exhibition days, the robot was running 9:20 hours and 6:30 hours, respectively, tracking about 2240 persons on the first day, and about 1400 persons on the second day. The large amount of persons tracked results from the following condition: Every person which came in the vicinity of the robot was counted once. However, if a person left the observed area and came back later, it was counted again as a new person.

Since the coffee breaks of the conference took place in the exhibition room, there were extremely busy phases. Even then, the robot was able to track up to 10 persons simultaneously. Despite the high noise level, the sound source localization worked reliably, even though it was necessary to talk slightly louder to attract the robot's attention.

8. SUMMARY

In this paper we presented a multi-modal attention system for a mobile robot. The system is able to observe several persons in the vicinity of the robot and to decide based on a combination of acoustic and visual cues whether one of these is willing to engage in a communication with the robot. This attentional behavior is realized by combining an approach for multi-modal person tracking with the localization of sound sources and the detection of head orientation derived from a face recognition system. Note that due to the integration of cues from multiple modalities it is possible to verify the position of a speech source in 3D space using a single pair of microphones only. Persons that are observed by the robot and are also talking are considered persons of interest. If a person of interest is also facing the robot it will become the current communication partner. Otherwise the robot assumes that the speech was addressed to another person present.

The performance of our approach and its robustness even in real world situations were demonstrated by quantitative evaluations in our lab and a qualitative evaluation during the exhibition of the mobile robot system at the ICVS'03.

9. ACKNOWLEDGMENTS

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'.

10. REFERENCES

[1] M. Andersson, A. Orebäck, M. Lindstrom, and H. I. Christensen. ISR: An intelligent service robot. In H. I. Christensen, H. Bunke, and H. Noltmeier, editors, *Sensor Based Intelligent Robots; International Workshop Dagstuhl Castle, Germany, September/October 1998, Selected Papers*, volume 1724 of *Lecture Notes in Computer Science*, pages 287–310. Springer, New York, 1999.

[2] B. Berdugo, J. Rosenhouse, and H. Azhari. Speakers' direction finding using estimated time delays in the frequency domain. *Signal Processing*, 82:19–30, 2002.

[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. Nat. Conf. on Artificial Intelligence (AAAI)*, pages 11–18, Madison, Wisconsin, 1998.

[4] S. Coradeschi and A. Saffiotti. Perceptual anchoring of symbols for action. In *Proc. Int. Conf. on Artificial Intelligence*, pages 407–412, Seattle, WA, 2001.

[5] Y. Freund and R. E. Shapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Computational Learning Theory: Eurocolt '95*, pages 23–27, 1995.

[6] J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Plötz, G. A. Fink, and G. Sagerer. Multi-modal anchoring for human-robot-interaction. *Robotics and Autonomous Systems, Special issue on Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems*, 43(2–3):133–147, 2003.

[7] J. Fritsch, S. Lang, M. Kleinehagenbrock, G. A. Fink, and G. Sagerer. Improving adaptive skin color segmentation by incorporating results from face detection. In *Proc. IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN)*, pages 337–343, Berlin, Germany, 2002.

[8] 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.

[9] E. Hjelmås and B. K. Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83(3):236–274, 2001.

[10] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau. Real-time passiv source localization: A practical linear-correction least-square approach. *IEEE Trans. on Speech and Audio Processing*, 9(8):943–956, 2001.

[11] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-24(4):320–327, 1976.

[12] 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, 2001.

[13] H. G. Okuno, K. Nakadai, and H. Kitano. Social interaction of humanoid robot based on audio-visual tracking. In *Proc. Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, Cairns, Australia, 2002.

[14] V. Pavlović, A. Garg, J. Rehg, and T. Huang. Multimodal speaker detection using error feedback dynamic bayesian networks. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, pages 34–43, Los Alamitos, CA, 2000.

[15] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[16] R. D. Schraft, B. Graf, A. Traub, and D. John. A mobile robot platform for assistance and entertainment. *Industrial Robot*, 28(1):29–34, 2001.

[17] D. Schulz, W. Burgard, D. Fox, and A. B. Cremers. Tracking multiple moving objects with a mobile robot. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 371–377, Kauwai, Hawaii, 2001.

[18] R. Stiefelhagen, J. Yang, and A. Waibel. Estimating focus of attention based on gaze and sound. In *Workshop on Perceptive User Interfaces*, Orlando, FL, 2001.

[19] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuro Science*, 3(1):71–86, 1991.

[20] P. Viola and M. Jones. Robust real-time object detection. In *Proc. IEEE Int. Workshop on Statistical and Computational Theories of Vision*, Vancouver, Canada, 2001.

[21] J. Yang and A. Waibel. A real-time face tracker. In *Proc. IEEE Workshop on Applications of Computer Vision*, pages 142–147, Sarasota, Florida, 1996.

[22] M. H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[23] A. L. Yuille, P. W. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *Int. Journal of Computer Vision*, 8(2):99–111, 1992.

[24] Z. Zhang, L. Zhu, S. Z. Li, and H. Zhang. Real-time multi-view face detection. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Washington, DC, 2002.