Proc. Autonomous Mobile Systems (AMS), Stuttgart, pp. 131-137, Informatik aktuell, Springer 2005

Integration of a Sound Source Detection into a Probabilistic-based Multimodal Approach for Person Detection and Tracking

Robert Brückmann, Andrea Scheidig[1], Christian Martin,
and Horst-Michael Gross

Ilmenau Technical University,
Department of Neuroinformatics and Cognitive Robotics

Abstract. Dealing with methods of Human-Robot-Interaction and using a real mobile robot, stable methods for people detection and tracking are fundamental features of such a system and require information from different sources. Based on an existing probability-based and multimodal approach for person detection and tracking, in this paper, we discuss the integration of a further sensory cue. This sensory cue is a sound source detection emerged from auditory information. Firstly, we discuss a newly developed approach for a sound source detection applied for a real world problem, dealing with the difficulty of reverberant environments. Secondly, we show a possible solution to integrate the sound source detection into the already existing person detection and tracking system applied for the mobile interaction robot HOROS working in a real office environment.

1 Introduction

Dealing with Human-Robot-Interaction (HRI) especially in real-world environments, one of the general tasks is the realization of a stable people detection and the respective tracking functions. Depending on the specific application that integrates a person detection, different approaches are possible. For real world problems, most promising approaches combine different sensory channels like visual cues and the scan of a laser-range-finder. Beside these sensory cues in the context of HRI also the auditory cue yields important information of the position of an interaction partner. Exemplary approaches which combine such different sensory cues are the SIG robot (auditory and visual cues) [7] or the BIRON project [2] (laser-range-finder, visual and auditory cues). The drawback of these approaches is the sequential processing of the sensory cues. For instance, people are detected by the laser information only and are subsequently verified by visual cues. Problems occur, when the laser-range-finder yields no information, for instance in situations when only the face of a person is perceivable.
To overcome this drawback, in [6] we propose a multimodal approach to realize the detection of people and the respective tracking functions. As sensory channels in [6] we use the following sensory modalities of our mobile interaction robot HOROS: the omnidirectional camera, the sonar sensors, and the laser-range-finder. A main advantage of our approach is the simple integration of further

sensory channels, like sound sources because of the used aggregation scheme. So in this paper, we firstly present an approach for sound source detection (see section 4), that will be integrated in the whole tracker system (see section 3). In result people can be detected by their legs, their faces and also by their speech based interaction or by only one of these features respectively. Respective results for the sound source detection in the context of a real world application will be shown in section 4.

2 Robot System HOROS

To investigate respective methods, we use the mobile interaction robot HOROS as an information system for employees, students and guests of our institute. The system's task includes that HOROS autonomously moves in the institute, detects persons as possible interaction partners and interacts with them, for example, to answer questions like the current whereabouts of specific persons.
The hardware platform for HOROS is a Pioneer-II-based robot from ActiveMedia. For the purpose of HRI, this platform was extended with different modalities. This includes a Tablet PC running under Windows XP for touch-based interaction, speech recognition and speech generation. It was further extended by a robot face which includes an omnidirectional fisheye camera, two webcams, and two microphones.
Laser-based Information: The laser-range-finder is a very precise sensor with a resolution of one degree, perceiving the frontal 180 degree field of HOROS. The laser-range-finder is fixed on the robot approximately 30 cm above the ground. Therefore it can only perceive the legs of people. Based on the approach presented in [1], we also analyze the scan of the laser-range-finder for leg-pairs using a heuristic method.
Sonar Information: Information from the sonar tends to be very noisy, imprecise und unreliable. Therefore, the variances are large and the impact on the certainty of a hypothesis is minimal. Nevertheless, the sonar is included to support people tracking behind the robot. So we are able to form an estimate of the distance in vision-based hypotheses.
Fisheye Camera: For HOROS we use an omnidirectional camera with a fisheye lens yielding a 360 degree view around the robot. Because of the task of person detection, the usage of such a camera requires that the position of the camera is lower than the position of the faces. To detect people in the omnidirectional camera image a skin-color-based multi-target-tracker[8] is used. This tracker is based on the condensation algorithm[3] which has been extended, so that the visual tracking of multiple people at the same time is now possible. A person detection using omnidirectional camera images yields hypotheses about the direction of a person but not about the distance.
Sound Source Detection: There are two electret-microphones attached to the head of HOROS which are used to detect acoustic sources. The distance between them is approximately 27 cm. With the detection algorithm described in section 4 the angle between the sound source and the robot can be calculated by

using the time delay of the sound. The possible resolution of the angle is up to two degrees for sources right in front of the robot. The two microphones don't allow for a full 360-degree-detection. Only sources between -90 and +90 degrees can be detected. Thus the combination with other sensory cues is necessary to avoid wrong detections.

The integration of the information from the camera and the sound source detection with the information from the laser-range-finder and the sonar sensors results in a powerful person detection system. Subsequently the developed method for the combination of the sensory systems will be discussed.

3 Generation of User Models

At first, a suitable data representation for the aggregation of the multimodal hypotheses resulting from the different sensor readings has to be choosen. The possibilities range from simple central point representation to probability distributions approximated by particles. The aggregation scheme we use is based on Gaussian distributions. Because of the unknown correlations between the different sensor readings, we did not use a Kalman Filter based approach to combine these hypotheses. Instead Covariance Intersection is applied [6].

First for the purpose of tracking, the sensory information about detected humans is converted into Gaussian distributions. The mean of each Gaussian distribution equals the position of the detection and the covariance matrix represents the uncertainty about this position. The form of the covariance matrix is sensor dependent due to different sensor characteristics, like their accuracy. Furthermore, the sensors have different error rates of misdetections that have to be taken into account.

Tracking based on probabilistic methods attempts to improve the estimate of the position of a human at each time. These estimates are integrated into a local map that contains all hypotheses around the robot. This map is also used to aggregate the sensor hypotheses from the current sensor readings. A sensor reading and a hypothesis with a minimum distance are merged. This update is done via the *Covariance Intersection* rule [4]. Sensor readings not matched with a hypothesis of the local map are introduced as a new hypothesis.

4 Integration of Sound Source Detection as a further Hypothesis

Besides the other sensory cues of the multimodal tracker, information gathered from sound sources is another important input for an interaction system. When implementing a sound source detection, especially reverberation often leads to wrong detection results. Therefore, we present a new approach for dealing with this issue.

Detection of sound sources can be achieved by using at least two spatially separated microphones. These receive the sound with a time delay which can be

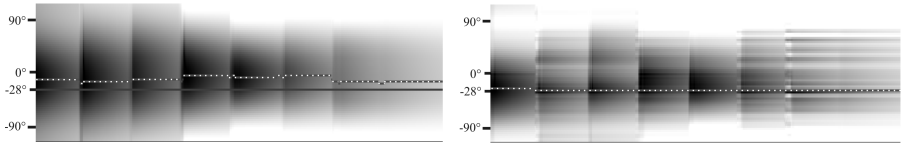

Fig. 1. Localization results compared to an unmodified cross-correlation for a speaker counting from 1 to 8 in a reverberant room. Distance to the robot was approx. 6 m, angle was approx. -28 degrees. **Left:** Results using only standard cross-correlation. The correct angle to be detected is marked with a dark-gray line at -28 degrees, the actual result is shown by the dotted line. **Right:** Results using the proposed localization approach. The dotted line shows the maximum of the correlation function at -28 degrees.

used to calculate the angle between the microphone array and the sound source. Hence our approach of detecting a speaker is based on the time delay of arrival (TDOA) between two microphones. In absence of noise and reverberation, the cross-correlation is a good method to measure the TDOA value:

$$r_{ij}(t) = \int_{-\infty}^{+\infty} x_i(\tau) x_j(t+\tau) d\tau \tag{1}$$

where x_i is the signal of microphone i. The position of the maximum in the cross-correlation represents the delay between the two signals. The cross power spectrum

$$R_{ij}(\omega) = X_i(\omega) \cdot X_j^*(\omega) \tag{2}$$

is the Fourier transform of the cross-correlation, where $X_i(\omega)$ is the Fourier transform of $x_i(t)$. Using the cross power spectrum, the influence of each frequency component can be weighted. The phase transform proposed in [5] uses the cross power spectrum enhanced by such a weighting function:

$$r_{ij}^{(g)}(t) = \int_{-\infty}^{\infty} \Psi_{ij}(\omega) X_i(\omega) X_j^*(\omega) e^{j\omega t} d\omega \tag{3}$$

where $\Psi_{ij}(\omega)$ is defined as

$$\Psi_{ij}(\omega) = \frac{1}{\left| X_i(\omega) X_j^*(\omega) \right|} \tag{4}$$

This weighting function normalizes the Fourier spectrum by setting the absolute values for all ω to 1. Thus only the phase of each frequency component remains. This whitening of the data narrows the resulting peak in the cross correlation function $r_{ij}^{(g)}(t)$ making the detection of the TDOA value easier.

A drawback of this transform is that every frequency bin of the Fourier spectrum will have the same influence to the resulting cross correlation, even if it is dominated by noise or contains reverberation. We added another weight to

the phase transform which provides the possibility to weight different frequencies according to their probability to contain reverberation.

$$\Psi_{ij}^{(e)}(\omega) = \frac{w(\omega)}{\left|X_i(\omega)X_j^*(\omega)\right|} \tag{5}$$

The function $w(\omega)$ can be used to decrease the influence of a frequency bin if it contains reverberation.

It is assumed that reverberation is received after the direct sound with a room-specific delay because the echo always has to cover a longer distance. By applying the cross-correlation only to the beginning of a perceived sound we can improve the results of the phase transform. A kind of onset-filter is used to implement this behaviour. The digital audio data of the two microphones is processed using windows of 1024 samples at a sample frequency of 44.1 kHz. Using the Fast Fourier Transform (FFT) we calculate the Fourier coefficients $X_i(k)$ of the windows for each microphone i. The weighting function for the discrete spectrum is expressed by

$$\Psi_{ij}^{(e)}(k) = \frac{w(k)}{\left|X_i(k)X_j^*(k)\right|} \tag{6}$$

We use thresholds for each frequency component to calculate the weights. The threshold values and the weights are adapted after the processing of each window.

$$w^t(k) = \min\left(0, X(k) - o^t(k)\right) \tag{7}$$

$$o^{t+1}(k) = \begin{cases} X^t(k) & , o^t(k) < X^t(k) \\ \alpha \cdot o^t(k) & , o^t(k) \geq X^t(k) \end{cases} \tag{8}$$

with $X^t(k)$ being the mean power spectral density of the two microphone channels. If a peak in one frequency bin has been found, then $X^t(k)$ is usually much larger than $o^t(k)$. The weight won't suppress the influence of this frequency for the current window. Subsequently this frequency band is inhibited for some time by raising the threshold $o(k)$. This ensures that the weight will be nearly 0 for the following windows. Therefore the reverb following the direct sound will not be evaluated in the cross correlation. It will decay over time and the threshold values will be decreased with each window using the decay factor $\alpha \in (0 \ldots 1)$. α determines how fast a frequency band will be available for the detection of new onsets. For our implementation, we empirically set α to 0.95. Such an onset detection is used separately for each frequency component. For speech signals, where the spectrum of the sound changes over time, it is possible to detect new onsets in different frequency bands while other bands are inhibited due to reverberation.

The angle between the microphone array and the sound source can be computed if the TDOA is known. We use a model to estimate this angle which assumes that the distance to the sound source is much larger than the distance between the two microphones.

$$\Theta = \arcsin \frac{v_{sound} \cdot \Delta t_{TDOA}}{\Delta s_m} \tag{9}$$

The distance between the two microphones is described by Δs_m, v_{sound} is the sound velocity.

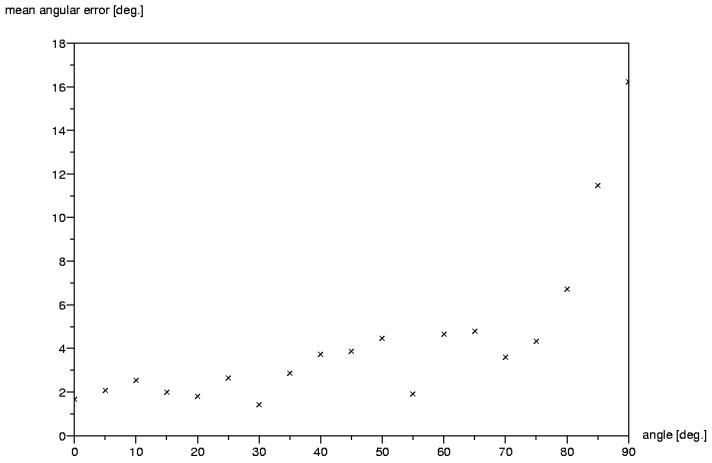

Fig. 2. Mean angular error of the sound source detection as a function of the actual angle between the robot and the sound source.

To test the localization algorithm, we placed a loudspeaker at different positions around the robot at a distance of approximately 75 cm and presented a recording of a male speaker. The mean angular error for different angles is shown in figure 2. Because of the non-linear relation between the TDOA and the computed angle, the resolution of the detection decreases for sources located on the sides of the robot. Practically, angles over 70 degrees only result in a rough direction estimation, whereas source right in front of the robot can be located quite accurately. Since our system is developed for Human-Robot-Interaction, the robot will turn towards the speaker. So the sound source will always be in front of the robot after the rotation, making the detection of the speaker easier.

Integration into a multimodal tracker

The integration of the found angle as a new hypothesis for the multimodal tracker is done by adding a Gaussian distribution representing the perceived speaker. It turned out that different types of audio signals yield different results with respect to localization robustness. A wide-band signal like a hand clap is generally easier to detect than a narrow-band signal, eg. a sine tone. Additionally, louder signals yield better localization results because of their higher signal-to-noise ratio. To take such details into account, the covariance of the Gaussian distribution illustrates the uncertainty depending on the broadness of the spectrum in conjunction with the overall loudness of the perceived audio signal. The

following uncertainty value can be used for the determination of the covariance:

$$Q_{angle} = \sum_{k=1}^{N} (w(k) \cdot X(k)) \tag{10}$$

This quality value increases if there are many frequency bins which contain loud signals and which are not inhibited by small weights. High values of Q_{angle} result in a narrow covariance while low values lead to wider covariances.

5 Summary and Conclusions

In this paper we have shown an approach for detecting a sound source using two microphones. A new method has been described allowing detection even in reverberant environments. The result has been integrated with a multimodal sensor tracker supporting the detection of people by adding a hypothesis of a possible speaker position. In our future work, we will adjust the balancing between the already existing sensory cues and the newly integrated tracker hypothesis generated by the sound source localization.

References

1. J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Ploetz, 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.
2. A. Haasch, S. Hohenner, S. Huewel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G.A. Fink, J. Fritsch, B. Wrede, and G. Sagerer. Biron - the bielefeld robot companion. In *International Workshop on Advances in Service Robots*, pages 898–906, May 2004.
3. M. Isard and A. Blake. Condensation - conditional density propagation for visual tracking. *International Journal on Computer Vision*, 29:5–28, 1998.
4. S. Julier and J. Uhlmann. A nondivergent estimation algorithm in the presence of unknown correlations. In *Proceedings of the 1997 American Control Conference*, pages 2369–2373 vol.4. IEEE, June 1997.
5. C.H. Knapp and C. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(4):320–327, 1976.
6. C. Martin, E. Schaffernicht, A. Scheidig, and H.-M. Gross. Sensor fusion using a probabilistic aggregation scheme for people detection and tracking. In *Proc. of ECMR*, 2005.
7. K. Nakadai, H.G. Okuno, and H. Kitano. Auditory fovea based speech separation and its application to dialog system. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2002)*, volume 2, pages 1320–1325, 2002.
8. T. Wilhelm, H.-J. Boehme, and H.-M. Gross. A multi-modal system for tracking and analyzing faces on a mobile robot. In *Robotics and Autonomous Systems*, volume 48, pages 31–40, 2004.