Eyes and Ears for a Humanoid Robot

D. Bechler, M. Schlosser, K. Kroschel
Institut für Nachrichtentechnik
Universität Karlsruhe
Kaiserstr. 12, 76128 Karlsruhe, Germany
{bechler,schlosser,kroschel}@int.uni-karlsruhe.de

R. Stiefelhagen[1], K. Nickel[1], A. Waibel[1,2]
Interactive Systems Laboratories
[1]Universität Karlsruhe
[2]Carnegie Mellon University, USA
stiefel@ira.uka.de, ahw@cs.cmu.edu

Abstract

For safe and efficient human-robot interaction, human friendly robots must have the perceptive capabilities to localize their users and to capture their communicative cues such as gestures or gaze direction. In this paper we present our ongoing work on building such perceptive components for a humanoid robot. First, we describe an attention system for the robot. Such a system is necessary to focus the robot's limited resources to the most important regions in the scene. We use a microphone array to accoustically track a user. We describe the design of the system, provide details of the used algorithms and present experimental results. For visual tracking of the user, a stereo camera is used. We track a user's face, hand and forearm locations in 3D by combining color and range information obtained from the stereo camera. An important cue for human-robot interaction is the user's focus of attention expressed by gaze direction. Here we present our approach to estimate a user's head pose from facial images using neural networks.

1 Introduction

In order to build collaborative human-friendly robots which are able to respond appropriately to their users' needs and to assure user safety, we need to equip these robots with the perceptive capabilities to capture all the necessary information about their users and the context in which they act. As an equivalent counterpart to the human's eyes and ears, robots use cameras and microphones as sensors for visual and acoustical perception.

Acoustically, the goal is to create an acoustic map of the robot's sound environment. For the acoustic scene analysis three main tasks have to be processed: localization, separation and classification of sound sources present in the acoustic scene. The main focus lies on the speech signal and its enhancement to guarantee a trouble-free speech-based interaction with humans through the audio channel. But also background sound signals are analyzed as they can be important indicators of salient objects and events or dangerous situations.

Visually, locating and capturing the user and his actions are of main interest. The robot should be able to detect and track persons that are nearby, it should be able to recognize their faces, monitor their gestures and body posture as well as their gaze.

Fusing acoustical and visual information is not only helpful for user tracking purposes but is also necessary for a complete analysis of the robot's environment. Therefore, the robot must be equipped with an attention system, detecting salient events and pointing out dangerous situations.

This paper is organized as follows: In Section 2 an attention system is presented dealing with an efficient analysis of the environment and the detection of dangerous situations to assure the user's safety. Section 3 describes our system to acoustically track a person with a microphone array. In Section 4 we describe our approach to track a person's face, hands and forearms in 3D using stereo vision. In Section 5 we discuss our approach to detect a person's focus of attention and provide details about our neural network based method to head pose estimation. Finally, some conclusions are drawn and an outlook on future work is given.

2 Attention System

A humanoid robot must be able to cooperate efficiently and safely with humans in an unconstrained environment. Therefore, the objectives of an attention system are to allow to learn efficiently about a new environment and to react to important events. Furthermore, the robot should not only present no danger to humans itself, but also be able to detect dangerous situations.

2.1 Efficiency and Data Reduction

It is computational prohibitive for a robot to process every region in every captured image to the highest cognitive levels, like object recognition and action planning. It has to bundle its limited resources on regions likely to contain important objects and events. As the human brain faces the same problem and per-

forms remarkably well on this task, it is sensible to use it as a starting point.

In the preattentive part of the human visual system several topographical feature maps like intensity, opponent-color, orientation, depth and movement are extracted in parallel over the complete visual field. These maps pass through a hierarchical structure with higher levels consisting of smaller maps and consistently larger receptive fields per pixel [1],[2].

On the one hand, this hierarchy provides scale invariance and, on the other hand, it calculates center-surround differences, which indicate salient objects, and performs the fusion into a single saliency map, which is used to determine the next point to focus attention on. Although skin color and faces are no features in the human system, it might be a good idea to include them into an artificial system as humans are of particular importance to the system.

2.2 Safety

As the capabilities of digital signal processing systems are still rather limited compared with those of humans, simple special mechanisms should be introduced to identify dangerous situations. To this end, a detector for objects moving at high velocity is presented in the followings. They not only represent an impending danger themselves naturally, but a quick movement of the user may also indicate a dangerous situation, which the robot has missed to realize itself.

Thus the objective is to find a computationally inexpensive algorithm that is nevertheless able to detect fast moving objects reliably. Our approach is based on optical flow images as such images are likely to be needed for other tasks anyway.

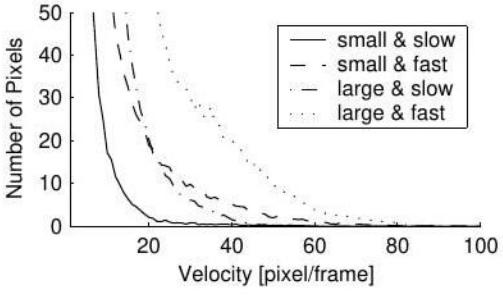

Figure 1: Averaged Velocity Histogram

Fig. 1 shows a typical distribution of velocity values in an optical flow image calculated using the algorithm in [3] averaged over 30 images. It can be approximated by a superposition of two independent exponential distributions. One for the static background and another one for the moving object. The former having a small decaying constant and the latter having a decaying constant proportional to the velocity and size of the moving object.

It can be seen that it is rather difficult to distinguish between a small object moving at high velocity and a large object moving at low velocity, a waving hand and a walking human respectively. At least two sampling points, one below and one above the intersection, need to be used.

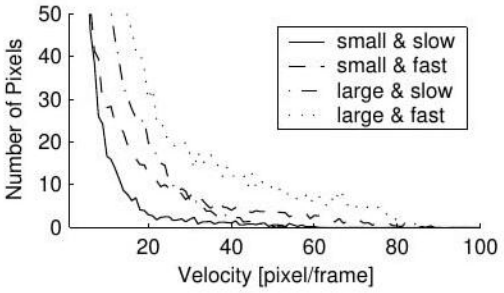

Figure 2: Standard Deviation of Velocity Histogram

The situation is aggravated further as the standard deviation of the velocity distribution is of the same order of magnitude as the velocity itself and thus the SNR is rather low, as can be seen in Fig. 2. To improve the SNR an integration operation, i. e. areas instead of points, should be used.

Their lower and upper bounds have to be chosen according to what velocities are regarded as indicating danger, and the capturing conditions. Furthermore, their choice is a trade-off between errors due to low SNRs and errors due to small relative differences between the different scenarios, i. e. small and large integration intervals respectively. On the other hand, the threshold values are a trade-off between false alarm and error rate.

Figure 3: Detected object moving at high velocity

If a velocity image is found to contain an object moving at high velocity, thresholding and morphological operations are applied to highlight this object, as can be seen in Fig. 3.

3 Acoustic Tracking

The technique of choice of most recent acoustic localization systems using microphone arrays is a two-step procedure. First, the time delay of arrival (TDOA) of speech signals in a pair of spatially separated microphones is estimated. In a second step the estimated TDOAs of different microphone pairs are used in combination with the microphone array geometry to localize the sound source.

Signal Model

For a given pair of spatially separated microphones M_i and M_j, the recorded sensor signals $x_i(t)$ and $x_j(t)$ for a signal $s(t)$, emanated from a remote sound source in a reverberant and noisy environment, can be modeled mathematically as

$$\begin{aligned} x_i(t) &= h_i(t) * s(t) + n_i(t) \\ x_j(t) &= h_j(t - \tau_{ij}) * s(t) + n_j(t), \end{aligned} \quad (1)$$

where τ_{ij} represents the relative signal delay of interest, $*$ signifies the convolution operator, $h_i(t)$ is the acoustic impulse response between the sound source and the i^{th} microphone and the additive term $n_i(t)$ summarizes the channel noise in the microphone system as well as environmental noise for the i^{th} sensor. This noise $n_i(t)$ is assumed to be uncorrelated with $s(t)$.

TDOA Estimation with GCC Method

The most popular approach for determining the TDOAs is called the Generalized Cross-Correlation (GCC) method [4]. The relative time delay τ_{ij} is estimated as the time lag with the global maximum peak in the GCC function $R_{ij}^{(g)}(\tau)$

$$\hat{\tau}_{ij} = \underset{\tau}{\operatorname{argmax}}\, R_{ij}^{(g)}(\tau). \quad (2)$$

This GCC function $R_{ij}^{(g)}(\tau)$ is defined as

$$R_{ij}^{(g)}(\tau) = \int_{-\infty}^{+\infty} \psi_{ij}(\omega) X_i(\omega) X_j(\omega)^* e^{j\omega\tau} d\omega. \quad (3)$$

The weighting function $\psi_{ij}(\omega)$ intends to decrease noise and reverberation influences and tries to emphasize the GCC value at the true TDOA value τ_{ij}. For real environments the Phase Transform (PHAT) technique [4] has shown best performance. This PHAT weighting function is defined as

$$\psi_{ij}^{PHAT}(\omega) = \frac{1}{|X_i(\omega) X_j(\omega)^*|}. \quad (4)$$

Confidence Criteria for TDOA estimates

For outlier detection for TDOA estimates, two confidence criteria can be used: the value of the maximum peak and the ratio between the 1^{st} and the 2^{nd} peak in the GCC function [5]. These criteria allow a reliability scoring of individual estimates and can be used to reject erroneous measurements. In combination with data association and clustering techniques the TDOA estimates are sufficiently accurate so that the following localization algorithm can produce robust sound position estimates.

Localization Algorithm

To come from the TDOAs and the microphone array geometry to the source position, the exact localization necessitates solving a set of non-linear equations, which can be computationally demanding. To accelerate the sound source position determination, the One-Step Least-Squares (OSLS) algorithm is used. This closed-form location estimator approximates sufficiently accurate the exact solution to the non-linear problem [6].

Post-Processing: Adaptive Kalman Filtering

For a continuous source trajectory, these initial, noise-corrupted position estimates are spatially smoothed by a Kalman Filter (KF). The applied KF is derived from 3 possible source motion models: a static, a constant velocity and a constant acceleration model. The motion dynamics of a speaker in our office environment can be variable and, hence, the decision for one of the sound source models has to be made continuously. Therefore, the approach of the Multiple Model Adaptive Estimator (MMAE) is applied [7]. In this approach 3 KFs with the 3 different motion models mentioned above run in parallel. The final position estimate is the sum of the estimates of each of these filters weighted with the according source motion model probabilities. These model probabilities can be calculated recursively from the initial input estimates.

Experiments and Results

Real data experiments have been carried out in a (5m x 5m x 3m) office room. For the data recording we used a 5-microphone array in an equilateral double-tetrahedron geometry with a side length of $D = 28$ cm as shown in Fig. 4. The sampling frequency was $f_s = 16$ kHz. The recorded data were analyzed in frames of 32 ms to assure quasi-stationarity. For this data segmentation a Hamming window with a 50% overlap was applied.

The proposed system shows robust speaker localization capabilities for our noisy and reverberant environment. The speaker could be tracked with ease if he or she keeps talking while moving. Exemplarily, Fig. 5 displays the true trajectory (arrow) and the positional estimates before and after the adaptive KF for a walking speaker in a 3D-plot. Note the advantage of the KF: the source trajectory is not only smoothed but also guaranteed to be continuous because of the

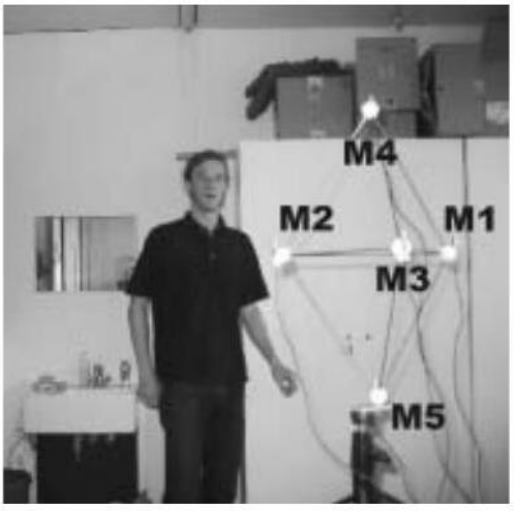

Figure 4: Experimental Setup

ability of the filter to predict source positions in case of missing current position estimates due to speech pauses or non-reliable TDOA estimates.

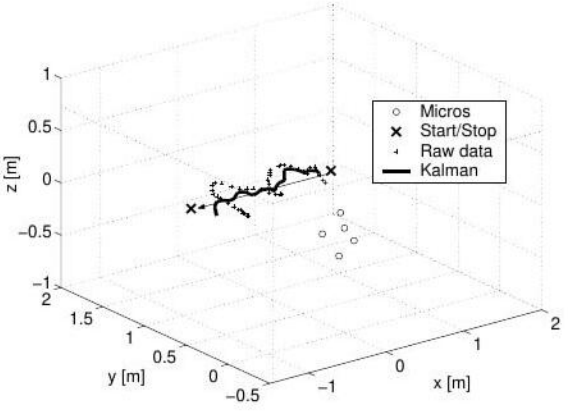

Figure 5: Positional Estimates before and after Adaptive Kalman Filtering

4 Visual Tracking

In order to gain information about the location and posture of a person interacting with a robot, we track the 3D-positions of the person's head and hands. They are important features for the recognition of many gestures, including the pointing gesture.

Our approach combines color information and range information obtained by stereoscopic vision, thus improving the quality of tracking compared to plain color-based tracking.

Vision System

Figure 6: Stereo camera head mounted on a pan-tilt unit

Our setup consists of a fixed-baseline stereo camera head mounted on a pan-tilt unit (Fig. 6).

A commercially available library (SRI's Small Vision System) calculates a dense disparity map made up of pixel-wise disparity values, and provides 3D-coordinates for each pixel. The pan-tilt unit is programmed to keep the head of the tracked person slightly above the center of the image.

Skin Color Modeling

Head and hands can be located by color as human skin color clusters in a small region of the chromatic color space [8]. To model the skin color distribution, two histograms of color values are built by counting pixel samples belonging to either the skin-color class S^+ or the *not*-skin-color class S^-. By means of the histograms, the ratio between $P(S^+|x)$ and $P(S^-|x)$ is calculated for each pixel x of the color image, resulting in a grey-scale map of skin-color probability.

To eliminate isolated pixels and to produce closed regions, a combination of morphological operations is applied to the skin-color map.

Combining Color and Range Information

In order to initialize and maintain the skin-color model automatically, we search for a person's head in the disparity map of each new frame. Following an approach proposed in [9], we first look for a human-sized connected region, and then check its topmost part for head-like dimensions. Pixels inside the head region contribute to S^+, all other pixels contribute to S^-. Thus, the skin-color model is continually updated to accommodate changing light conditions.

In order to find potential candidates for the coordinates of head and hands, we search for connected regions in the thresholded skin-color map. For each region, we calculate the spatial mean of the associated

3D-pixels, weighted by their skin-color probability. If the pixels belonging to one region vary strongly with respect to their distance to the camera, the region is split by applying a k-means clustering method. We thereby separate objects that are lying on different range levels, but accidentally merge into one object in the 2D-image.

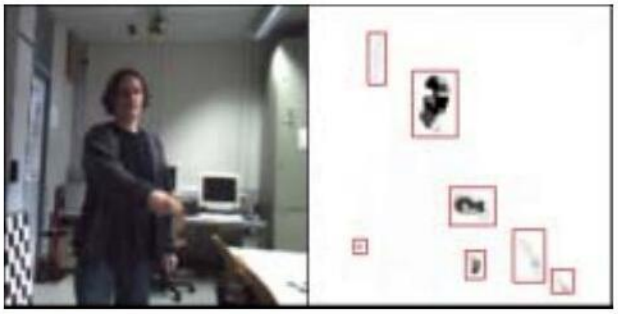

Figure 7: Head and hand candidates detected in the skin-color map

Fig. 7 shows the skin-color map generated for a video frame. Dark pixels represent high skin-color probability. All connected regions are highlighted, their spatial means are the potential coordinates of head and hands.

Tracking

The *state* of the tracker s_t is a combination of the 3D-coordinates of head and hands at time t. With each new frame, all permutations of the candidate coordinates are evaluated in order to find a new combination s_t that maximizes the product of following 3 scores:

The *observation score* $P(O_t|s_t)$ is a measure for the extent to which a given state s_t matches the observation O_t. $P(O_t|s_t)$ increases with each pixel that complies with the model, e.g. a pixel showing strong skin-color at a position the model suggests to be part of the head.

The *posture score* $P(s_t)$ is the prior probability of the posture. It is high if the posture represented by s_t is a frequently occurring posture of a human body. It is equal to zero if s_t represents a posture that breaks anatomical constraints. To be able to calculate $P(s_t)$, a model of the human body was built from training data. The model consists of the average height of the head above the floor, a probability distribution (represented by a mixture of gaussians) of hand-positions relative to the head, as well as a series of constraints like the maximum distance between head and hand.

The *transition score* $P(s_t|s_{t-1})$ is a measure for the probability of s_t being the successor of s_{t-1}. It is higher the closer the positions of head and hands in s_t are to their positions in the previous state s_{t-1}. $P(s_t|s_{t-1})$ is set to a value close to zero[1] if the distance of a body part between $t-1$ and t exceeds the limit of a natural motion within the short time between two frames.

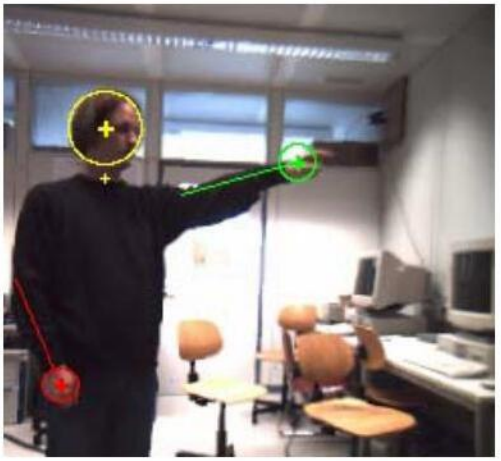

Figure 8: Automatically extracted 3D body features

Experimental Results

Our experiments indicate that by using the method described, it is possible to track a person robustly, even when the camera is moving and when the background is cluttered.

The tracking of the hands is affected by occasional dropouts and misclassifications. Reasons for this can be temporary occlusions of a hand, a high variance in the visual appearance of hands and the comparatively high speed with which people move their hands.

The detection of the head in the disparity image does not succeed in every frame, but it is still sufficient to keep the skin-color model up and running.

The system runs at about 8 frames per second on a standard 1GHz-PC.

Pointing Gestures

Pointing gestures are amongst the most important gestures a mobile robot should be able to understand. One approach to extract the direction of a pointing gesture is based on the assumption that the pointing direction is the extension of the line of sight between the head and the pointing hand [10]. As this approach is limited to gestures performed with an outstretched arm, our estimate of the pointing direction is based on the direction of the *forearm*. To identify the orientation of the forearm, we calculate the covariance matrix C of the 3D-pixels $x_{1..N}$ within a 20cm radius

[1] $P(s_t|s_{t-1})$ must always be positive, so that the tracker can recover from erroneous static positions.

around the center of the hand μ:

$$C = \frac{1}{N} \sum_{N} (x_n - \mu)(x_n - \mu)^T \quad (5)$$

The first principal component of C indicates the direction of the largest variance of the pixel positions, which is equivalent to the orientation of the forearm on condition that no other objects are present in the critical radius around the hand.

This work is in progress - an evaluation will have to show the accuracy of our approach. Fig. 8 shows a frame from a video sequence including the extracted locations of head and hands as well as the orientation of the forearms.

5 Focus of Attention

Gaze plays an important role in human social interaction. During face-to-face communication people look at each other, monitor each other's lip-movements and facial expressions, and follow each other's gaze. In an intelligent working space, where humans and robots interact with each other, information about the user's gaze direction is a necessary cue to detect with what or whom the user is interacting or to what he is paying attention.

In recent years, we have addressed the problem of tracking the visual focus of attention of participants in meetings; i.e., tracking who is looking at whom during meetings [11, 12]. In the framework of the Sonderforschungsbereich "Humanoide Roboter" we are now adapting and extending our approach to build a gaze-aware robot which is able to monitor a person's focus of attention.

A body of research literature suggests that humans are generally interested in what they look at (e.g. [13]) and the close relationship between gaze and attention during social interaction has been emphasized [14, 15]. In addition, recent user studies reported strong evidence that people naturally look at the objects or devices with which they interact [16, 17].

A first step to determine someone's focus of attention, therefore is, to find out in which direction the person looks. There are two contributing factors in the formation of where a person looks: head orientation and eye orientation. In this work head orientation is considered as a sufficient cue to detect a person's direction of attention. Relevant psychological literature offers a number of convincing arguments for this approach (e.g. [15, 14]) and it has been shown experimentally, for example, that head orientation alone is a very reliable cue to detect focus of attention of participants in a meeting [18].

A practical reason to use head orientation to estimate a person's focus of attention, is, that in scenarios such as addressed in this work, head orientation can be estimated with non-intrusive methods while eye orientation can not.

In the remainder of this section we describe our approach to estimate head orientation using neural networks.

5.1 Estimating Head Pose Using Neural Nets

In this work we aim at estimating head orientation directly from facial images. The main advantage of such an appearance based approach is that no facial landmark points have to be detected in order to compute head pose. Instead, head pose is estimated from the whole facial image and therefore only the face has to be detected and tracked in the camera image.

We use neural networks to estimate pan and tilt of a person's head from pre-processed facial images. Similar approaches are for example described in [19] or [20]. As compared to our work, these systems are, however, user-dependent and report only results for one single user. They also differ in the used network architectures and image preprocessing approaches.

Data Collection

We collected training data from 19 persons in our lab. During data collection, users had to wear a head band with a sensor of a Polhemus pose tracker attached to it. Using the pose tracker, the head pose with respect to a magnetic transmitter could be collected in real-time. Figure 9 shows two sample images that were taken during data collection.

Figure 9: Two good resolution images taken with a pan-tilt-zoom camera during data collection.

Preprocessing of Images

To locate and extract the faces from the collected images, we use a statistical skin color model [8]. The largest skin colored region in the input image is selected as the face.

Two different image preprocessing methods were investigated: 1) Using normalized grayscale images of the user's face as input and 2) applying edge detection to the images before feeding them into the nets.

In the first preprocessing approach, histogram normalization is applied to the grayscale face images as a means towards normalizing against different lighting conditions. No additional feature extraction is performed. The normalized grayscale images are downsampled to a fixed size of 20x30 pixels and are then used as input to the nets.

In the second approach, a horizontal and a vertical edge operator plus thresholding is applied to the facial grayscale images. The resulting edge images are downsampled to 20x30 pixels and are both used as input to the neural nets.

Figure 10 shows the corresponding preprocessed facial images of a user. From left to right, the normalized grayscale image, the horizontal and vertical edge images of a user's face are depicted.

Figure 10: Preprocessed images: normalized grayscale, horizontal edge and vertical edge image (from left to right)

Neural Net Architecture, Training and Results

We have trained separate nets to estimate head pan and tilt. For each net, a multi-layer perceptron architecture with one output unit (for pan or tilt) and one hidden layer with 20 to 150 hidden units was used. We estimate head pan in the range of -90 to +90 degrees and head tilt in the range of -60 to +60 degrees. Output activations for pan and tilt were normalized to the range [0,1]. Training of the neural net was done using standard back-propagation.

We used around 9900 images from 17 different users to train a "multi-user" network. Both the cross-evaluation set and the test-set contained around 1240 images from the same seventeen users. After training, we achieved a mean error of 3.8 degrees for pan and 3.2 degrees for tilt on the test set.

To determine how well the neural net based system can generalize to new users, we have also evaluated the neural networks for pan and tilt estimation on data from two new users, which have not been in the training set. On these new users an average error of 7.1 degrees for pan and 9.5 degrees for tilt was obtained.

Table 1 summarizes the results on the multi-user test set and on the new users.

Test Set	E_{pan}	E_{tilt}
multi-user	3.8	3.2
new users	7.1	9.5

Table 1: Average estimation error in degrees for pan and tilt on a multi-user test set and on two new users.

5.2 From Head Pose to Focus of Attention

Once a person's head orientation is estimated, we would like to infer the most likely target – such as objects or persons in the scene, or the robot itself – at which a person might have looked at. In our previous work on focus of attention tracking in meetings [11, 12] we have developed a statistical approach to find the most likely target *person* at which a subject might have looked at, based on his or her head orientation.

In our proposed approach the class-conditional head orientation distrubitions for the detected target persons are modelled as Gaussians and are automatically learned by looking at a subject's head orientations over time. The current approach is however limited to scenarios, where the number of participants around a table and their locations remains the same within a meeting.

We are currently investigating how to extend our approach to make it work in a more dynamic environment where as well the user, the robot and the objects in the scene may move.

6 Conclusions

In this work some key components for a humanoid robot to be able to share a common environment with a human were presented. These consist of a attention system to reduce the data stream and detect dangerous situations, acoustic and visual systems to locate a user and to learn about the user's intentions. For the latter, systems to determine the user's focus of attention and his pointing gestures were presented. They naturally combine with speech recognition to a multimodal communications interface.

The work is still in progress. Further extensions are envisioned to meet the demands of our scenario. Already mentioned were the robustness of the recognition of the pointing gestures and the adaptation of the target models to dynamic environments for the estimation of the user's focus of attention.

As far as the attention system is concerned a special mechanism to detect objects falling to the ground is planned as this is again likely to indicate a dangerous situation. Furthermore, a acoustical component shall be incorporated serving as an early-warning system to indicate events outside the visual field of the cameras.

The acoustic tracking system shall be extended

to multi-source environments with background sound sources. Furthermore, it is planned to perform a complete acoustic scene analysis including the separation and classification of these sound sources

We also plan to combine accoustic and visual person tracking in order to get a more robust person tracker. Work is in progress to combine them by means of a Kalman filter. In designing this filter, special emphasis should be placed on the modeling of the measurement errors. To this end, possible confidence measures were already presented in Section 3.

7 Acknowledgments

This work is part of the Sonderforschungsbereich (SFB) No. 588 *"Humanoide Roboter - Lernende und kooperierende multimodale Roboter"* at the University of Karlsruhe. The SFB is supported by the Deutsche Forschungsgemeinschaft (DFG).

References

[1] L. Itti. *Models of Bottom-Up and Top-Down Visual Attention.* PhD thesis, California Institute of Technology, Jan 2000.

[2] J. M. Wolfe. *Attention*, chapter Visual Search. London: UCL Press, 1996.

[3] B.K.P. Horn and B.G. Schunck. Determining optical flow. A. I. Memo 572, Massachusetts Institute of Technology, 1980.

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

[5] D. Bechler and K. Kroschel. Confidence scoring of time difference of arrival estimation for speaker localization with microphone arrays. In *13. Konferenz Elektronische Sprachsignalverarbeitung ESSV*, September 2002.

[6] Y. Huang, J. Benesty, and G.W. Elko. Passive acoustic source localization for video camera steering. In *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages 909–912, June 2000.

[7] Y. Bar-Shalom. *Tracking and data association.* Academic Press, 1988.

[8] Jie Yang and Alex Waibel. A real-time face tracker. In *Proceedings of WACV*, pages 142–147, 1996.

[9] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. In *IEEE Conference on Computer Vision and Pattern Recognition, (Santa Barbara, CA)*, pages 601–608, June 1998.

[10] N. Jojic, B. Brumitt, B. Meyers, S. Harris, and T. Huang. Detection and estimation of pointing gestures in dense disparity maps. In *4th IEEE International Conference on Face and Gesture Recognition*, pages 468–475, March 2000.

[11] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13(4):928–938, July 2002.

[12] Rainer Stiefelhagen. Tracking focus of attention in meetings. In *International Conference on Multimodal Interfaces*, pages 273–280, Pittsburgh,PA, October 2002. IEEE.

[13] A. L. Yarbus. Eye movements during perception of complex objects. In L.A. Riggs, editor, *Eye Movements and Vision*, pages 171–196. Plenum Press, New York, 1967.

[14] Michael Argyle and Mark Cook. *Gaze and Mutual Gaze.* Cambridge University Press, 1976.

[15] N.J. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*, 24:581–604, 2000.

[16] Paul P. Maglio, Teenie Matlock, Christopher S. Campbell, Shumin Zhai, and Barton A. Smith. Gaze and speech in attentive user interfaces. In *Proceedings of the International Conference on Multimodal Interfaces*, volume 1948 of *LNCS*. Springer, 2000.

[17] B. Brumitt, J. Krumm, B. Meyers, and S. Shafer. Let there be light: Comparing interfaces for homes of the future. *IEEE Personal Communications*, August 2000.

[18] Rainer Stiefelhagen and Jie Zhu. Head orientation and gaze direction in meetings. In *Conference on Human Factors in Computing Systems (CHI2002)*, Minneapolis, April 2002.

[19] Bernt Schiele and Alex Waibel. Gaze tracking based on face-color. In *International Workshop on Automatic Face- and Gesture-Recognition*, pages 344–348, 1995.

[20] Robert Rae and Helge J. Ritter. Recognition of human head orientation based on artificial neural networks. *IEEE Transactions on neural networks*, 9(2):257–265, March 1998.