Robotics and Autonomous Systems, 44 (2003), pp.83-96, Elsevier Science

An Approach to Multi-modal Human-Machine Interaction for Intelligent Service Robots

Hans-Joachim Böhme [a,3] Torsten Wilhelm [a] Jürgen Key [a,1] Carsten Schauer [a,2] Christof Schröter [a,1] Horst-Michael Groß [a] Torsten Hempel [b]

[a] *Ilmenau Technical University, P.O.B. 100565, 98684 Ilmenau, Germany*

[b] *LGI Logistic Group International, Hewlett-Packard Street 2, 71034 Böblingen, Germany*

Abstract

The paper describes a multi-modal scheme for human-robot interaction suited for a wide range of intelligent service robot applications. Operating in un-engineered, cluttered, and crowded environments, such robots have to be able to actively contact potential users in their surroundings and to offer their services in an appropriate manner. Starting from a real application scenario, the usage of a robot as mobile information kiosk in a home store, some reliable methods for vision-based interaction, sound analysis and speech output have been developed. These methods are integrated into a prototypical interaction cycle that can be assumed as a general approach to human-machine interaction. Experimental results demonstrate the strengths and weaknesses of the proposed methods.

Key words: Human-Robot Interaction, People Detection, People Tracking, Service Robot Application, Multi-modal Interaction

1 Introduction

Intelligent service robots, a research field that became more and more popular over the last years, cover a wide range of application scenarios, from robotic assistance for disabled or elderly people up to climbing machines for cleaning large storefronts. Our specific scenario is aimed at the development of an intelligent interactive shopping assistant, working as a mobile information kiosk in a home store (see fig. 1). In contrast to the application of personalized robots, where robot and

[1] Partially supported by Thuringian Ministry of Science, Research and Arts
[2] Partially supported by Deutsche Forschungsgemeinschaft, Graduiertenkolleg GK164
[3] Corresponding author, e-mail: hans-joachim.boehme@tu-ilmenau.de

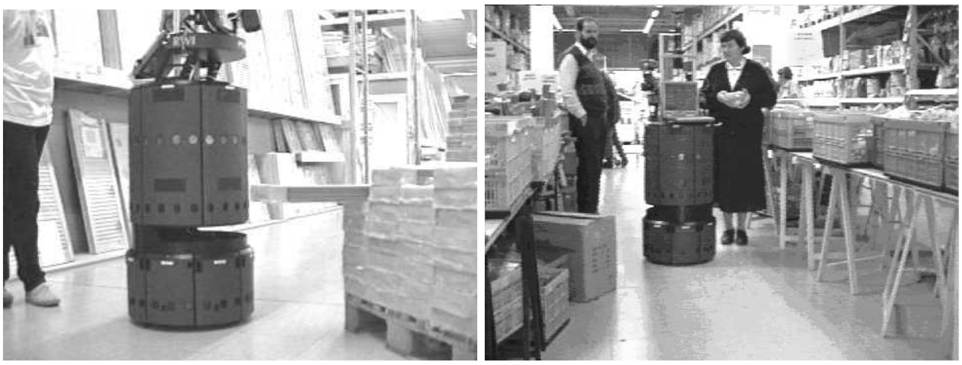

Fig. 1. Our experimental platform PERSES operating in a home store, a cluttered and un-engineered environment.

user can adopt to each other, such a robot has to be able to interact with anybody. Furthermore, these people typically know neither the scope of the robot nor its functional capabilities. People have no idea of how the robot works, if it has a name by which it may be called, or if it understands speech at all. In general, for robots working in public places, an intuitive interactive behavior is a necessary prerequisite for the acceptance of such robots by their potential users. When looking at stationary information terminals often placed in shopping centers, these terminals are almost always an eyesore. One major reason for that fact is that these terminals are not interactive in a natural sense. They cannot detect if there is anybody interested in the information provided, but repeat their information repertoire endlessly. To preserve service robots from the same fate, we suppose that a natural, intuitively understandable interaction scheme is urgently needed. Such an interaction scheme should contain components everybody is familiar with, during everyday human-to-human interaction. Consequently, vision and acoustics should play the major role.

During the past decade, a variety of approaches to intelligent human-robot interfaces has been proposed ([3, 15, 7]). Most of them argue, as we do, that the combined utilization of speech and vision channel seems the most appropriate way for building such interfaces.

As stated above, we are particularly interested in a more general framework, whereas most of the previous approaches are very specific for a certain domain.

First of all we want to summarize typical behavioral skills of an interactive service robot. The system has to contact potential users in its surroundings, to verify if the person is interested, to offer its services, and finally to keep continuous contact during the whole interaction process. This collection takes into account the necessities of our application scenario, but the mentioned skills are valid for service robot applications in general.

In our proposed interaction scheme, the first step contains the generation of hypotheses concerning people in the surroundings of the robot and is called *person detection* throughout the paper. Here, a vision-based movement detection and an analysis of acoustic signals are combined into an attentional process, that results in a turning of the robot towards the most salient direction. Then, a *person localization* procedure rechecks if there really is a person and if the person could be interested

in using the robot. For the case that an interested person approaches the robot, the robot welcomes and offers its services. This is realized by means of situation dependent speech output and a graphical user interface running on a touch-screen. As long as the current user remains in the (visible) surroundings of the robot, the robot tries to keep continuous contact to its user via *person tracking*.
The remainder of the paper is structured as follows. After introducing the robot and its technical setup, section 2 describes the developed methods in detail. In section 3, experimental results are given and an exemplary interaction process is demonstrated. Section 4 contains ongoing and complementary work as well as some summarizing conclusions.

2 Methods for Multi-modal Human-Robot Interaction

2.1 The Robot PERSES

Fig. 2 shows the robot PERSES, an extended version of a standard mobile robot B21 by RWI (IS Robotics). In addition to the common equipment of two sonar and one IR-layers, PERSES utilizes (i) an omnidirectional color camera with a 360^o panoramic view used for user localization and tracking, self localization and local navigation, (ii) a binocular 6 DoF active-vision head with 2 frontally aligned color cameras used for user localization and tracking, odometry correction and obstacle avoidance, (iii) a binaural auditory system for acoustic user localization and tracking, and (iv) a touch-screen for immediate user-robot interaction.

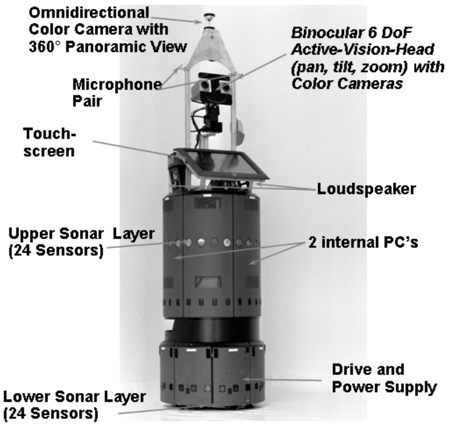
Omnidirectional Color Camera with 360° Panoramic View
Binocular 6 DoF Active-Vision-Head (pan, tilt, zoom) with Color Cameras
Microphone Pair
Touch-screen
Loudspeaker
Upper Sonar Layer (24 Sensors)
2 internal PC's
Drive and Power Supply
Lower Sonar Layer (24 Sensors)

Fig. 2. Experimental platform PERSES.

2.2 Movement Detection within the Omnidirectional Images

For every mobile service robot, one major problem consists in the robust detection and localization of a potential user in its operation area. Our vision-based user detection performs a motion-based foreground-background segmentation in the input images provided by the omnidirectional camera. In the waiting position or while standing still, the motion-based segmentation provides some candidate regions that indicate if and where people could be in the surroundings of the robot. The implemented method is similar to that suggested in the *Pfinder* system [21], but differs in the following aspects: (i) the statistical models for foreground and background pixels were simplified to boxes, and (ii) the foreground and background models are adjusted to the current situation. The model simplification led to a lower computational load resulting in a performance speed-up, surprisingly almost without any loss in sensitivity. By adaptation of the foreground and background models, we take into account that the robot cruises its surroundings which makes it impossible to use only one stationary background model, but in fact, the method presupposes that the robot does not move itself during movement detection. After the alignment of all image pixels to the foreground and background model, respectively, some appropriate heuristics are used to assess the motion for every angular direction. These heuristics are needed to determine what direction the most attractive one could be. The corresponding assessment parameter relies on three different aspects: (i) The direction of motion indicates, if the person is moving towards the robot or not, and a person moving away from the robot or passing the robot is probably no candidate for interaction. (ii) The size of the moving regions gives information concerning the distance of that object (person) to the robot. The lower the size of the moving region the larger the distance between object and robot can be assumed. (iii) The angle difference between the robot's current orientation and the direction(s) where motion is detected gives a measure of how long the robot will take turning to that direction. For the case that several people surround the robot, this distance should be rather small, leading to fast turns to the nearest standing (moving) person. The implementation of those heuristics leads to the following behavior: the robot preferably turns towards people that are moving towards the robot and that are relatively close to the robot.

2.3 Sound Localization

For the acoustic detection of a potential user clapping her hands or shouting a command, we developed a biologically inspired model of binaural sound localization using inter-aural time differences and spikes as temporal coding principle [13]. This subsystem realizes (i) the detection of the sound direction in the horizontal half-planes by processing the inter-aural time-delays and (ii) a simple but effective front-behind discrimination on the basis of the differences in the spectral shapes of the left and right sound stream supplied by the microphones mounted on top of PERSES (Fig. 1). It detects pitch onsets in the signals and calculates the angle to the sound source from the phase shift between the binaural signals. Details of this model and localization results are presented in [16].

2.4 Fusion of Motion Detection and Sound Localization

The integration of auditory saliency makes it easy for the user to attract the attention of the robot to accelerate the localization process significantly. Both methods supply an angle by which the robot has to be turned. In case both angles drive the robot to the same direction, that direction is strongly supported. Otherwise, motion detection and sound localization work autonomously. In case motion detection and sound localization indicate different directions the robot will choose that one that would cause the lower turning movement. Consequently, a potential user can attract the robot's attention via ego-motion or, alternatively, by emitting a sound.

2.5 Person Localization

To evaluate if there really is a person and if she could be willing to interact with the robot, we developed a localization system that integrates different visual cues. This system should highlight the regions that most likely cover the upper part of a person. Concentrating on the upper part of a person has the following reasons: One has less difficulties concerning (partial) occlusions, and the features described below are very person-specific and indicate if the person is roughly aligned towards the robot. Execution of person localization is triggered, when the robot was turned by the person detection module. Fig. 3 gives an overview of the corresponding architecture. Due to the turn of the robot, the potential customer should be localized in front of the robot, allowing to observe her by the frontally aligned cameras as well as by the omnidirectional camera. Because we want to localize people even at different distances from the robot, a multi-resolution pyramid (scale space with five fine-to-coarse resolutions) transforms the images into a multi-scale representation. Two cue modules sensitive to *facial structure* and *structure of a head-shoulder contour*, respectively, operate at all levels of the gray-scale pyramid. The cue module for *skin color* detection uses the original color image. After superposition of the corresponding feature maps, a 3D-Winner-Take-All process within the saliency pyramid selects that region most likely covering the upper part of a person.
The utility of the different parallel processing cue modules is to make the localization system robust and independent of the presence of one certain information source in the images. Hence, we can more easily handle varying environmental circumstances that, for instance, make the skin color detection difficult or almost impossible.

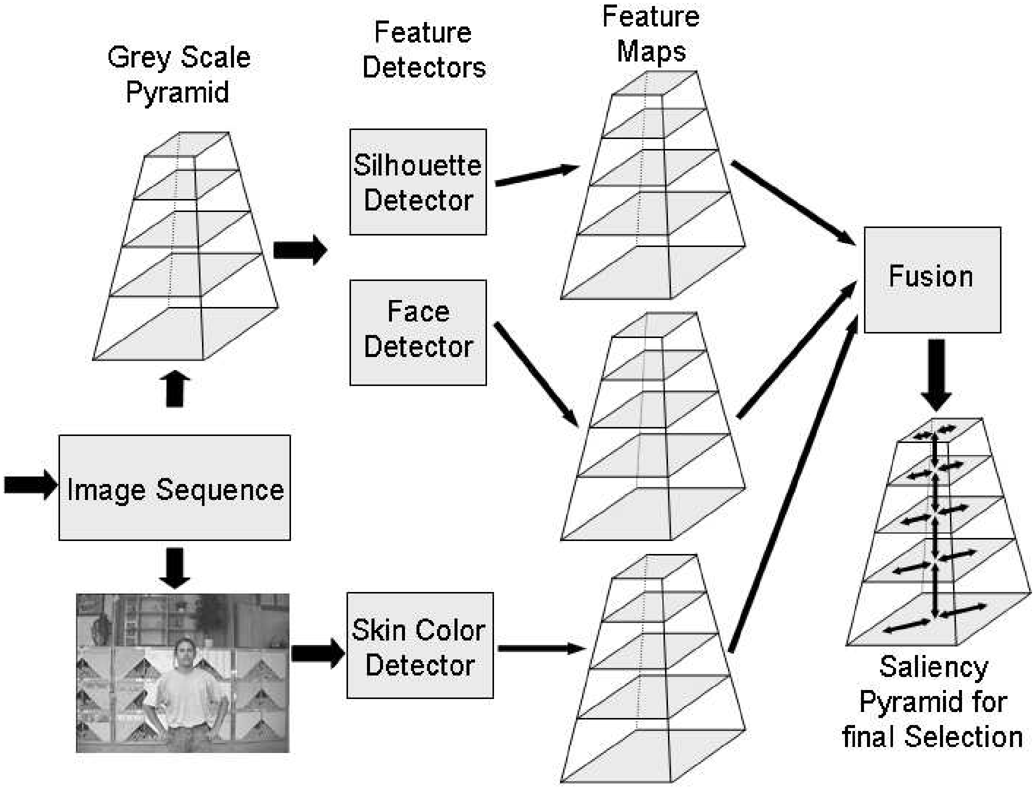

Fig. 3. Multiple-cue approach to user localization.

For person localization and the subsequent tracking process, both camera systems (omnidirectional as well as frontally aligned stereo system) are utilized.

Contour Modelling: The contour which we refer to is that of the upper part of the body of a frontally aligned person. First, we generated a statistically determined average head-shoulder contour by collecting views of different people. Then, a model was learned based on this set of training images (see fig. 4). This simple contour shape prototype consists of a course of orientations along the modelled contour and realizes a piecewise approximation of the upper shape of a person (head, shoulder). Applying such a contour model in a multi-resolutional manner leads to a robust localization of frontally aligned people even in depth. For computing the orientation

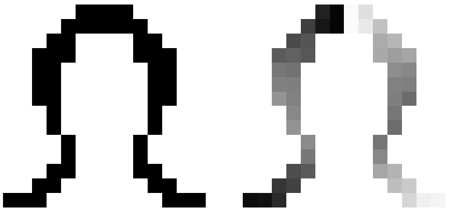

Fig. 4. Illustration of the statistically determined contour model by the binary contour shape (left) and the local orientation values along the contour (right). Orientation angles are coded by gray values (0^o: black; 90^o: medium gray; 180^o: white).

along the contour, a tensor-based method proposed in [11, 2] was implemented.

Compared to classical orientation-specific filtering with Gabor-wavelets [12] or steerable filters [9], this method is faster by orders of magnitude. Orientation filtering provides a pair containing the dominant orientation angle and the strength of the contour at that point. The bandpass dimension determines the extent of the local area where the orientation is calculated. By varying the dimension of the applied bandpass filters it is possible to create a feature jet for each pixel. The components (tuples) of such a jet code different dominant orientations, dependent on the applied bandpass filter. For contour localization, we utilize a specific distance measure taking into account the difference between extracted and expected orientation value at every contour point as well as the contribution of each contour point to the whole contour model. Furthermore, the discontinuity between 0 and 180 degree (angle wraparound) is handled by doubling the angle within distance calculation. Distance measure and jet representation allow a two-step coarse-to-fine search in orientation space (see fig. 5). First, a pre-selection is done via a coarse distance threshold and the orientation values obtained with one bandpass dimension (two-dimensional manifold of orientation space) resulting in a few candidate contour locations. Then, these preselected candidates are finally checked using the whole orientation space. This procedure is much less time consuming compared to applying the fine search for every image location.

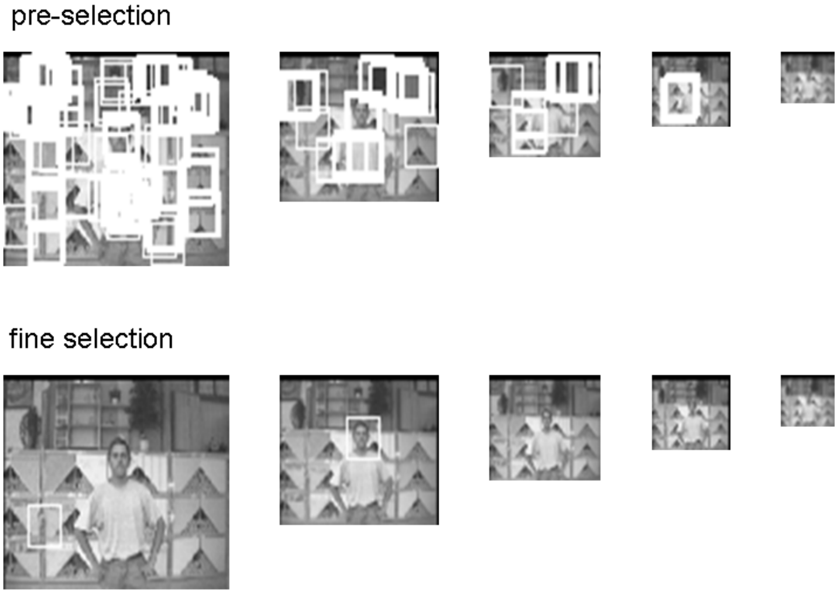

Fig. 5. Illustration of the two-step coarse-to-fine contour localization method. The upper row contains the results of the coarse contour localization (pre-selection), whereas the bottom row depicts the remaining contour hypotheses after the fine analysis using the whole orientation space. The shown five layers of the scale space cover a distance range from 0.5 up to about 3 meters.

Skin Color Detection: Skin color is a typical feature for person detection and person tracking. Usually, a color space is employed where color and intensity information are uncorrelated. A widely used skin color modelling procedure was suggested in [22] and is also applied in our system. A set of skin colored pixels was generated by acquiring images of different people (skin types) under varying lighting conditions (illumination colors). This data collection is transformed from the RGB color space into the dichromatic r-g color space and subsequently modelled by a bivariate normal distribution (fig. 6).

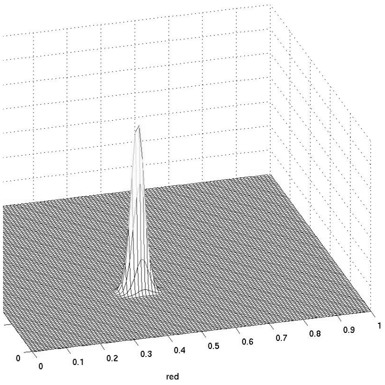

Fig. 6. Skin color distribution in r-g color space.

For skin color detection, the Mahalanobis-distance between the color values of a pixel and this model distribution gives us the likelihood for being skin colored (see the raw skin color classification in fig. 7b). To get closed skin colored regions, a median filter is applied at every resolution level of the scale space, followed by a segmentation algorithm.
Unfortunately, skin is not the only skin colored object. Therefore, some heuristics have been developed to improve the separation between real skin color and other skin colored image regions. For every resolution level the size of the skin colored regions as well as their width-height relation is checked according to the expected face region. Subsequently, regions that do not fit the applied criteria can be rejected. Fig. 7 depicts an example for the described skin-color processing regime.

Face Detection: Several approaches to face detection have been described, ranging from using Eigenfaces [19], feature based [23, 6] and neural network based methods [14]. The advantages of applying neural networks for the face detection task are quite obvious: The facial image is characterized directly in terms of pixel intensities, and according to the two-class problem at hand (face, no face) a training pattern set can be used to adjust the parameters of the classifier. But, training a neural network for face detection is challenging because of the difficulty in characterizing prototypical ”non-face” images. As suggested in [14], one can avoid this problem by using a bootstrap algorithm that adds automatically false positive classified image regions to the training pattern set as the training process progresses. The module for face detection is implemented as a Cascade-Correlation Neural Network (CCNNW, [5]). The reason for using that kind of neural network lies in its capability to produce a network topology that fits optimally with the complexity of the mapping problem. In contrast to the standard Multilayer Perceptron, where the

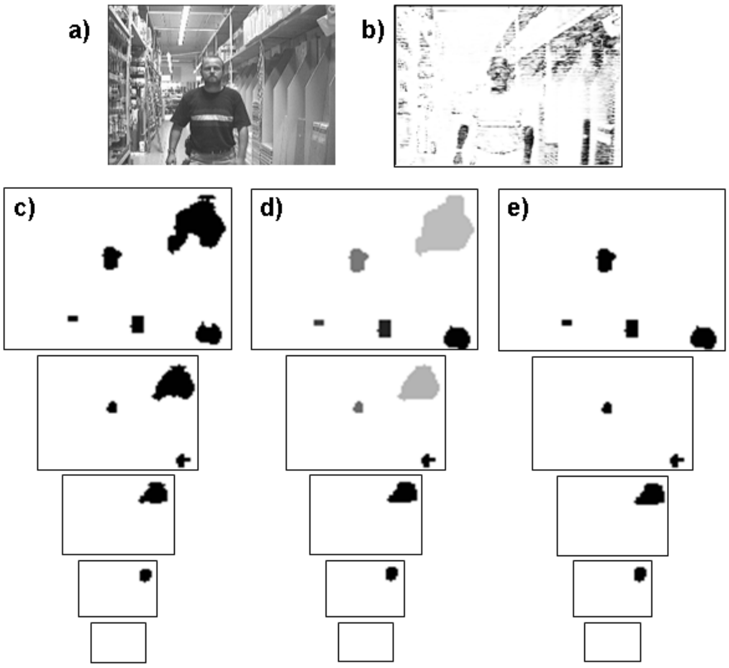

Fig. 7. Processing steps for skin color detection: a) original image, b) raw skin color classification, c) smoothing by applying of a median filter for all resolution levels, d) result of the segmentation algorithm, and e) final detection result according to the chosen heuristics for every resolution level.

network topology has to be chosen in advance (see [14]), the CCNNW optimizes the network parameters along with its topology during the same training process. Starting with a minimal topology (direct linear input-output mapping), new hidden nodes are trained to maximally reduce the networks output error, until a chosen termination criterion is fulfilled. Fig. 8 depicts the finally obtained topology for the CCNNW.

To generate a training pattern set for faces, 174 images out of a public data base provided by AT&T Laboratories Cambridge (http://www.cam-orl.co.uk/facetatabase.html) were utilized. From these images, 15×20 pixel sized regions covering only the face were manually extracted. Initially, the non-face pattern set contains a collection of randomly chosen images, and is extended during bootstrapping. An exemplary result obtained with the CCNNW-face detector is shown in fig. 9. Surprisingly, the face detector performs quite well even on the polar-cartesian transformed omnidirectional images, were in contrast to the training patterns local distortions of the face region occur, but this is only to demonstrate the generalization abilities of the face detector. A particular training of the face detector with face images coming from the omnidirectional camera has not been done because at the moment person localization is exclusively realized via the frontally aligned cameras.

Cue Fusion and Final Selection: The final step for obtaining the image region(s) that most likely cover the upper part of frontally aligned people contains a simple

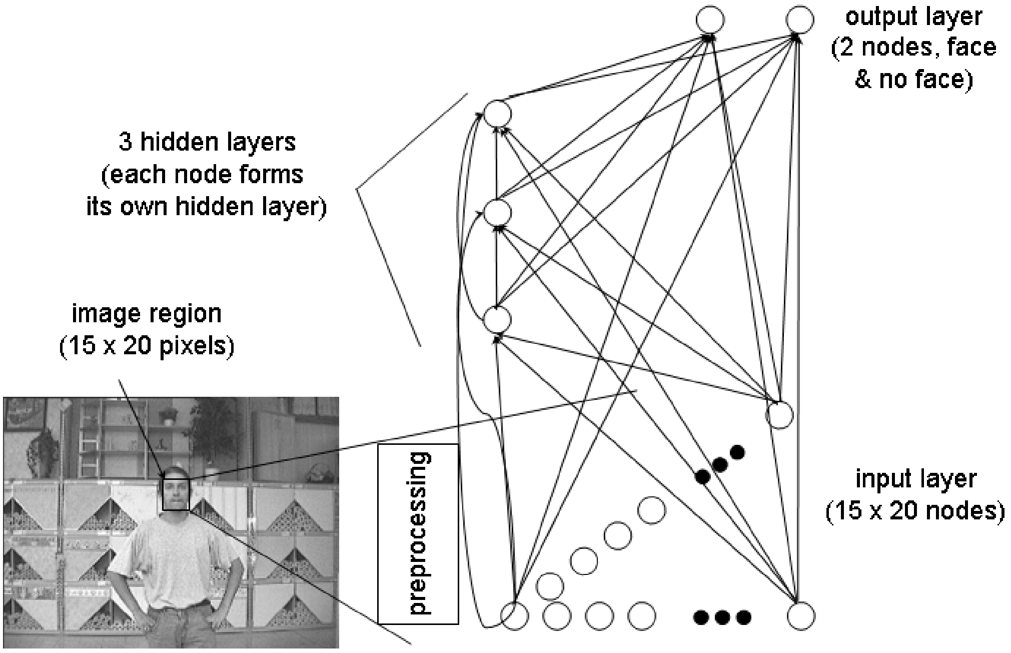

Fig. 8. Topology of the CCNNW used for face detection.

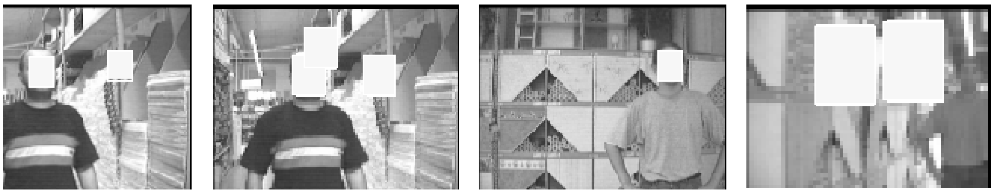

Fig. 9. Exemplary results for face detection with the Cascade-Correlation Neural Network, in front of highly cluttered background. From left to right: two images containing correct and false positive detections, an image with only correct detection, and an image where the face detection failed (only false positive detections).

fusion method and a subsequent selection mechanism. Only those image locations where at least two out of the three cues supply a detection result, are allowed to contribute to the final selection process (see fig. 10). To ensure that all cues are equally weighted during the selection process, a uniform Gauss-shaped activity blob is used to encode every detection result (image location).

The final selection process is realized by means of a dynamic neural field [1]. Since dynamic neural fields are powerful tools for dynamic selection using simple homogeneous internal interaction rules, we adapted them for our purposes. Because we use five fine-to-coarse resolutions in our scale space (see fig. 3), we can actually localize people even at different distances. Therefore, a neural field for selecting the most salient region should be three-dimensional. The field is described as a recur-

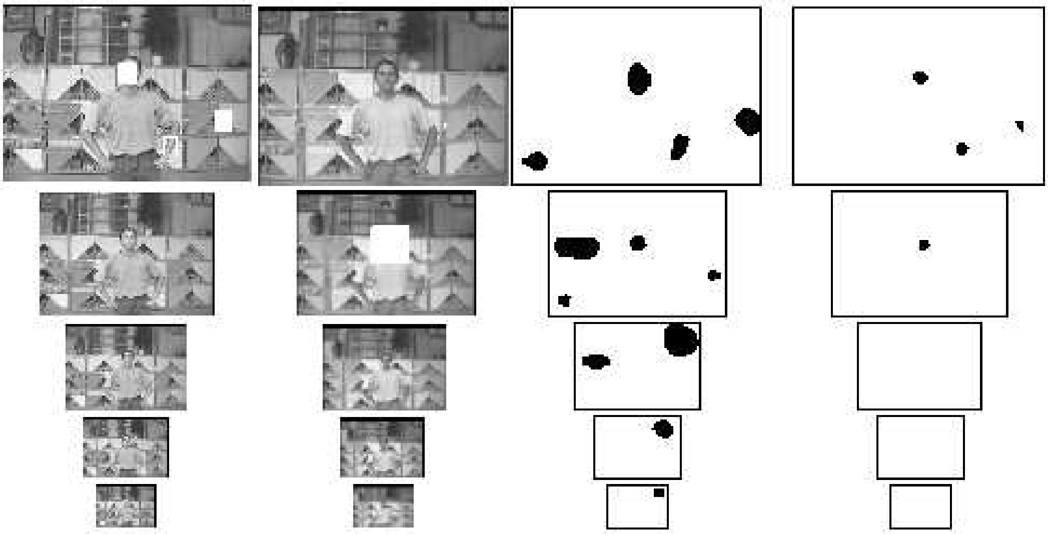

Fig. 10. One example for fusing the different cues for person localization. According to the fusion rule, only those locations of the scale space are fed into the final selection process were at least two cues supply a detection result. (From left to right): face detection, head-shoulder contour detection, skin color detection, and fusion result.

rent nonlinear dynamic system. Regarding the selection task, we need a dynamic behavior which leads to *one* local region of active neurons successfully competing against the others, i. e. the formation of one single blob of active neurons as an equilibrium state of the field (for a detailed description see [4]).

By using a three-dimensional neural field, we are able to consider the local correspondences within as well as between adjacent resolution levels. This leads to an interesting side effect: because outputs of the different cue detectors often occur at the same location of adjacent resolution levels, such correspondences enhance the selection of such locations, resulting in a much more robust localization.

2.6 Multi-modal People Tracking

The goal of people tracking is to keep continuous contact to the current user, and person localization provides the initialization for the subsequent tracking process. The general tracking procedure is based on the *Condensation* algorithm [10], widely accepted as a powerful and efficient method for tracking arbitrarily shaped probability distributions [8].

In principal, visual tracking can be done via the omnidirectional camera as well as via the frontally aligned cameras, but currently, only the omnidirectional camera is used. The features underlying this part of the tracking process were derived from the presented person localization procedure. A combination of head-shoulder contour detection and skin color modelling turned out to be appropriate. Both cues are subsumed by a Fuzzy-Minimum-Maximum operator. The output of this operator takes into account that both cues have to be present at corresponding image

locations up to a certain degree and determines the ”visual” weight of the samples within the Condensation algorithm.
Facing the complex environmental conditions in the home store, the purely visual tracking procedure suffers from highly variable lighting and scene background, resulting in a non-satisfying robustness. Therefore, people tracking is extended into a multi-modal approach combining visual and sonar information. Within the sonar scan we assess the distance to the person localized via the method described above. By continuously re-checking this distance measure against the visual cues silhouette and color, it is possible to re-weight the condensation samples. The integration of visual and sonar information leads to a very reliable people tracking method. A more detailed description of the implemented tracking algorithm can be found in [20].

2.7 Graphical User Interface, Speech Output and Robotic Face

Via the graphical user interface, running on a touch-screen that is mounted on top of the robot, an immediate interaction between robot and human user can be realized. In our application scenario, the customer can choose an item she is looking for or a desired market area. Generally, this kind of ”classical” human-machine interaction cannot be completely replaced in the near future. The reason is quite obvious: the appropriate alternative would be a purely speech-based dialog between robot and human, but, up to now, speech recognition methods do not possess the necessary capabilities concerning vocabulary size, associative mapping, context dependency, dialect and so on. Moreover, for service robots interacting with anybody, one cannot assume that robot and human operate within the same reference frame. In other words, the robot does not know what the human will say, and on the other side, the human has no idea about the vocabulary the robot is able to recognize.

In our opinion, speech output is much more than only entertainment. Via speech, the robot can tell its current state, can offer its services, or can ask its current user to solve ambiguous or uncertain situations. For simplicity reasons, we currently use prepared sound files, and their activation is triggered by certain situations. For instance, after successful person localization the robot welcomes this potential user and invites her to interact via the touch-screen.

Inspired by the smart face of MINERVA, the robotic tour-guide described in [18], PERSES was equipped with its own face, created by eye-like camera fronts as well as mouth and eyebrows made of controllable diode arrays (see fig. 11). Hence, the current ”emotional” state of the robot can be transmitted in a more natural and intuitive way.
Fig. 12 is to resume the overall interaction schema described so far. It contains the different processing steps and clarifies the information flow throughout the system.

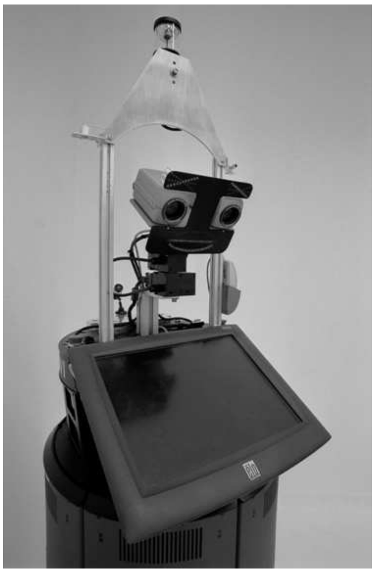

Fig. 11. Face of PERSES.

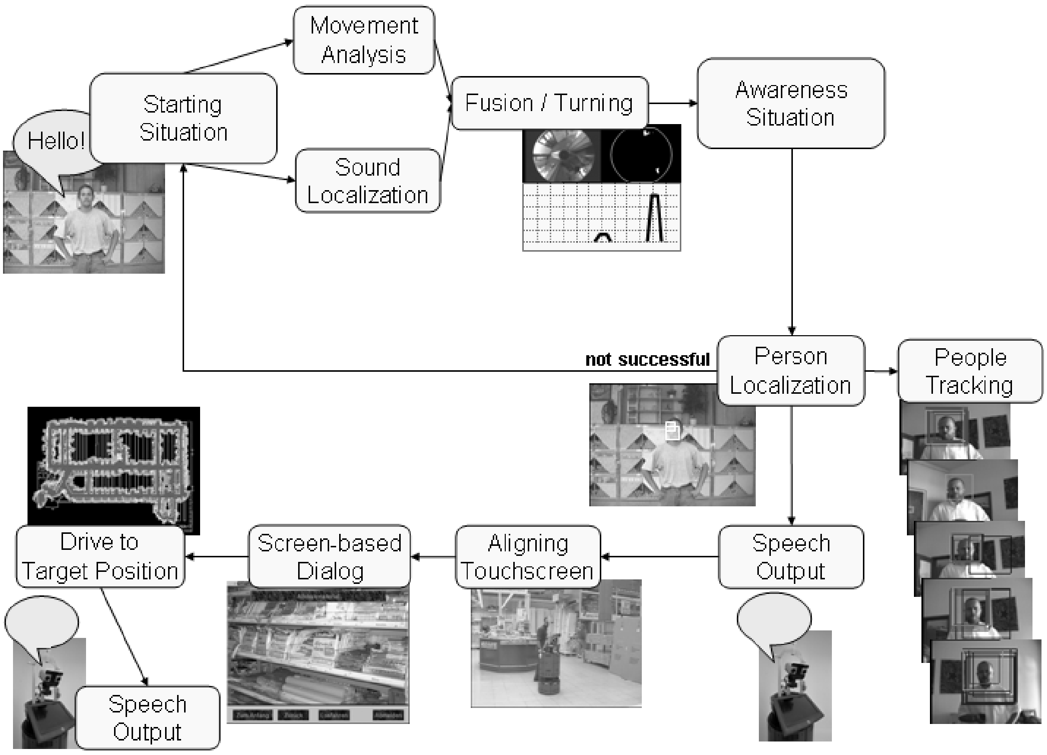

Fig. 12. Overall architecture of the proposed multi-modal interaction schema.

3 Experimental Results

The experiments shown below are to demonstrate an exemplary interaction cycle between service robot and its user in the home store.

Interaction starts with person detection. Within the omnidirectional view (top left in fig. 13) person P2 is moving towards the robot, whereas person P1 passes the robot. Both people are detected via motion-based segmentation (top right). Subsequently, the robot estimates the most attractive direction by valuation of the two different directions (bottom of fig. 13), and turns towards person P2.

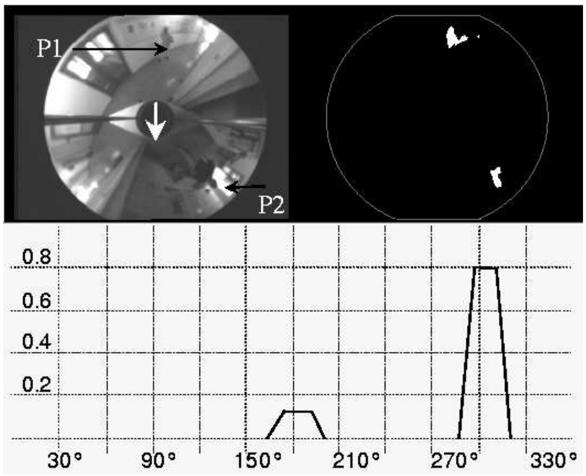

Fig. 13. Person localization via motion detection. The white arrow marks the view direction (direction of the frontally aligned cameras, 0^o) of the robot, the angle runs clockwise.

Person detection is followed by person localization. Fig. 14 contains a collection of localization results. The localization module provides an output only when cue fusion and final selection (see section 2.5) supply a very strong result. To avoid false localizations is very important, because otherwise the robot would start interacting with uninterested people or even with inanimate items. In case a potential customer has not been successfully localized, the customer can log-in directly via the touch-screen. When successful localization happened, the robot welcomes the customer by means of a typical speech sequence. Then, the customer can choose the desired item or the interesting market area via the touch-screen. After selection, the robot confirms the corresponding item and shows a map of the market, where its current position and the goal position are indicated.

During the whole interaction cycle, the robot tries to keep continuous contact to the current client via visual tracking. An exemplary tracking sequence is given in fig. 15, where samples of a longer run of the tracking system (over several minutes) are shown. Both sequences clarify the advantage of the multi-modal (visual / sonar) approach. By using the distance measure towards the people being tracked, the tracking algorithm can handle the crossing of the people within the surroundings of the robot without loosing the person in front of the other one. Furthermore, we avoid the utilization of a specific motion model for the tracked object. Such a motion model could also be used to distinguish between different moving people in the surroundings of the robot, but, unfortunately, it is very difficult to describe typical movements of people.

Although robot navigation behavior was not explicitly described, it should be noticed that people tracking and navigation have to be combined appropriately. For

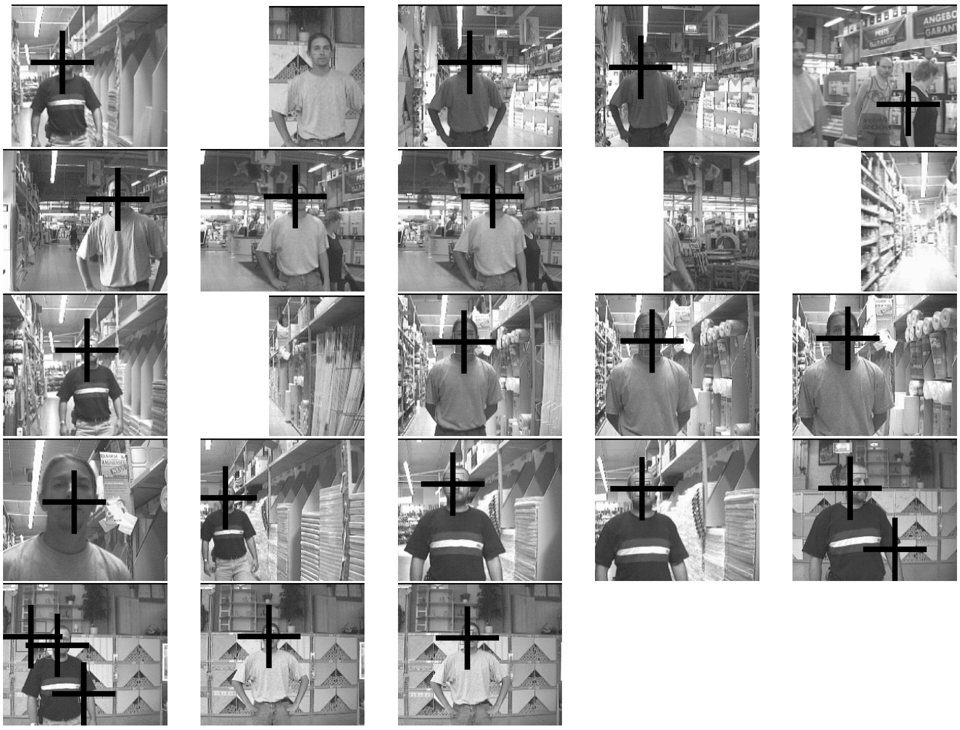

Fig. 14. Person localization results (black crosses) for different situations in the real home store. People occur in front of highly cluttered background. Special emphasis has been made to improve robustness, specificity and efficiency of the localization procedure. Currently, the system runs with 0.5 Hz on a Double-Pentium III (500 MHz) with an image resolution of about 200 × 200 pixels (frontally aligned camera, depth range from 0.5 up to 2.5 meters).

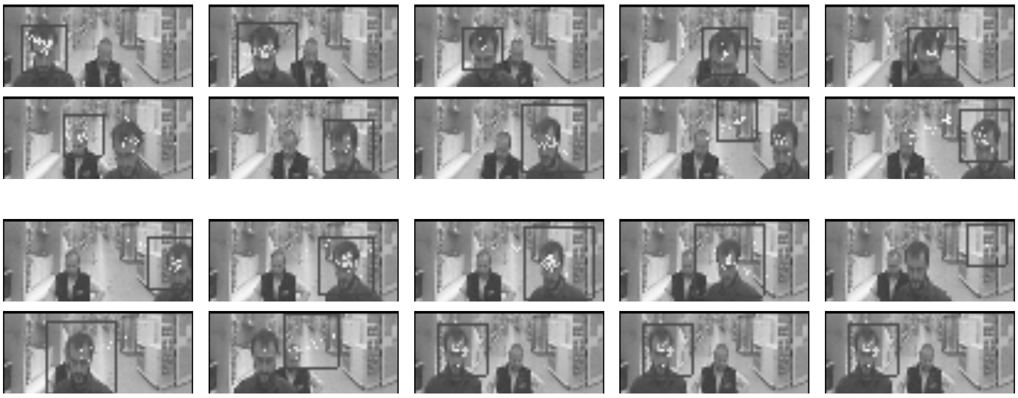

Fig. 15. People tracking experiments. Within the shown sequences (to read from left to right), the frames in the images of the omnidirectional camera indicate those locations where the tracked person is most likely expected.

example, when the robot tries to reach a predefined target location in the home store, direction and speed of the movement are adjusted according to the current tracking status. That means that the robot slows down when the distance to the client becomes to large. As long as the contact to the current user can be continuously updated, no articulation of the robot is needed. If the robot detects a situation where the user is lost, the robot stops and provides a speech output to ask the user to reduce the distance to the robot. Alternatively, the guidance to a desired market position is temporarily interrupted and the robot moves towards its present user to prevent losing contact.

4 Conclusions and Outlook

The paper has described a multi-modal scheme for intelligent and natural human-robot interaction. Special emphasis was placed on vision-based methods for user localization, person localization and person tracking and their embodiment into a multi-modal overall interaction schema. The proposed interaction regime should be understood as work in progress, undergoing continuous changes. The experimental results demonstrate the principal functionality of the corresponding subsystems. Although the interaction cycle strongly relates to our home store scenario, it can contribute to a wider range of service robot applications.

Future research will concentrate on the extension of the tracking system, which is currently limited to roughly frontally aligned people. This includes the vision-based methods as well as the integration of scan-based person tracking techniques as proposed in [17]. Furthermore, we will work on the design and implementation of a framework for modelling the interaction cycle to provide the robot with the capability to learn and generalize from a series of interactions with different people.

References

[1] Amari, S. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87, 1977.

[2] Bigun, J. and Granlund, G.H. Optimal orientation detection of linear symmetry. In *First International Conference on Computer Vision (ICCV)*, pages 433–438. IEEE Computer Society Press, 1987.

[3] Bischoff, R. and Graefe, V. Integrating Vision, Touch and Natural Language in the Control of a Situation-Oriented Behavior-Based Humanoid Robot. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume II, pages 999–1004, 1999.

[4] Boehme, H.-J., Brakensiek, A., Braumann, U.-D., Krabbes, M., and Gross, H.-M. Neural Architecture for Gesture-Based Human-Machine-Interaction. In *Gesture and Sign-Language in Human-Computer Interaction*, Lecture Notes in Artificial Intelligence, pages 219–232. Springer, 1998.

[5] Fahlman, S.E. and Lebiere, Ch. The cascade-correlation learning architecture. In *Advances in Neural Information Processing Systems 2*, pages 524–532. Morgan Kaufmann Publishers, Inc., 1990.

[6] Feyrer, S. and Zell, A. Tracking and Pursuing Persons with a Mobile Robot. In *International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems (RATFG-RTS '99)*, pages 83–88, 1999.

[7] Feyrer, S. and Zell, A. Robust Real-Time Pursuit of Persons with a Mobile Robot Using Multisensor Fusion. In *6th International Conference on Intelligent Autonomous Systems (IAS-6)*, pages 710–715, 2000.

[8] Fox, D., Delleart, F., Burgard, W., and Thrun, S. Monte Carlo Localization: Efficient Position Estimation for Mobile Robots. In *Proceedings 16th National Conference on Artificial Intelligence (AAAI-99)*, 1999.

[9] Freeman, W.T. and Adelson, E.H. The Design and Use of Steerable Filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 13(9):891–906, 1991.

[10] Isard, M. and Blake, A. CONDENSATION – conditional density propagation for visual tracking. *International Journal on Computer Vision*, 29(1):5–28, 1998.

[11] Jaehne, B. *Practical Handbook on Image Processing for Scientific Applications*. CRC Press LLC, 1997.

[12] Jones, J.P. and Palmer, L.A. An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex. *Journal of Neurophysiology*, 56(8):1233–1258, 1987.

[13] Paschke, P. and Schauer, C. A spike-based model of binaural sound localization. *International Journal of Neural Systems*, 9(5):447–452, 1999.

[14] Rowley, H. A., Baluja, S., and Kanade, T. Neural Network-Based Face Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[15] Roy, D. and Pentland, A. Multimodal Adaptive Interfaces. In *AAAI Spring Symposium on Intelligent Environments*, 1998. TR 438, M.I.T. Media Lab Perceptual Computing Section.

[16] Schauer, C., Zahn, T., Paschke, P., and Gross, H.-M. Binaural sound localization in an Artificial Neural Network. In *Proceedings IEEE-ICASSP'2000*, volume II, pages 865–868. IEEE Press, 2000.

[17] Schulz, D., Burgard, W., and Cremers, A.B. State Estimation Techniques for 3D Visualizations of Web-based Tele-operated Mobile Robots. *Künstliche Intelligenz*, 4:16–22, 2000.

[18] Thrun, S., Beetz, M., Bennewitz, M., Burgard, W., Cremers, A.B., Dallaert, F., Fox, D., Hähnel, D., Rosenberg, C., Roy, N., Schulte, J., and Schulz, D. Probabilistic algorithms and the interactive museum tour-guide robot minerva. *International Journal of Robotics Research*, 19(11):972–999, 2000.

[19] Turk, M. and Pentland, A. Face recognition using eigenfaces. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1991.

[20] Wilhelm, T., Böhme, H.-J., and Groß, H.-M. Sensor Fusion for Visual and Sonar based People Tracking on a mobile Service Robot. In *Dynamic Perception*, Proceedings in Artificial Intelligence. infix Verlag, 2002. to appear.

[21] Wren, C., Azarbayejani, A.and Darrell, T., and Pentland, A. Pfinder: Real-Time Tracking of the Human Body. *IEEE Transactions on Pattern Analysis*

and Machine Intelligence (PAMI), 19(7):780–785, 1997. M.I.T. Media Lab Techreport TR 353.

[22] Yang, J., Lu, W., and Waibel, A. Skin-Color Modeling and Adaptation. In *Computer Vision - ACCV*, volume 2, pages 687–694, 1997. CMU-CS-97-146, Carnegie Mellon University.

[23] Yow, K.C. and Cipolla, R. Feature-based Human Face Detection. *Image and Vision Computing*, 15:713–735, 1997.