Fourth European Workshop pm Advanced Mobile Robots (Eurobot 2001), pp. 17-24, Lund, Sweden, Lund University Cognitive Studies, vol. 86, 2001

AN APPROACH TO MULTIMODAL HUMAN-MACHINE INTERACTION FOR INTELLIGENT SERVICE ROBOTS

Hans-Joachim Boehme, Torsten Wilhelm, Juergen Key, Christof Schroeter, Horst-Michael Gross
Ilmenau Technical University, Ilmenau, Germany, hans@informatik.tu-ilmenau.de

Torsten Hempel
LGI, Boeblingen, Germany, torsten.hempel@lgi.de

Abstract
The paper describes a multimodal 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, noise 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.

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 assistent, working as a mobile information kiosk in a home store (see fig. 1). In contrast to the application of personalized robots, where robot and 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 centres, 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.

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

During the past decade, a variety of approaches to intelligent human-robot interfaces has been proposed ([2, 14, 6]). 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. Fig. 2 summarizes typical service tasks and behavioral skills of an interactive service robot. The sketch takes into account the necessities of our application scenario, but the mentioned skills are valid for service robot applications in general. The system has to contact potential users in its surroundings, to verify if the person is interested, to offer its services, and

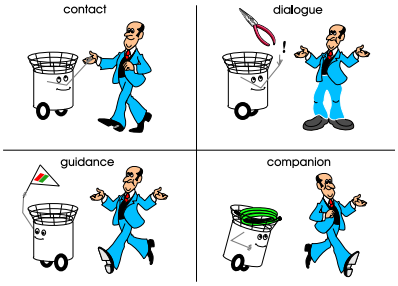

Figure 2: Necessary skills and typical service tasks of an interactive mobile robot (shopping assistent).

finally to keep continuous contact during the whole interaction process.

In our proposed interaction scheme, the first step contains the generation of hypotheses concerning people in the surroundings of the robot. 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 verification 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 grafical 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 examplary interaction process is demonstrated. Section 4 contains ongoing and complementary work as well as some summarizing conclusions.

2. Methods for Multimodal Human-Robot Interaction

2.1 The Robot Perses

Fig. 3 shows the robot Perses, an extended version of a standard mobile robot B21 by RWI (IS Robotics). In addition to the standard equipment of two sonar and one IR-layers, Perses is equipped with (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 verification 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.

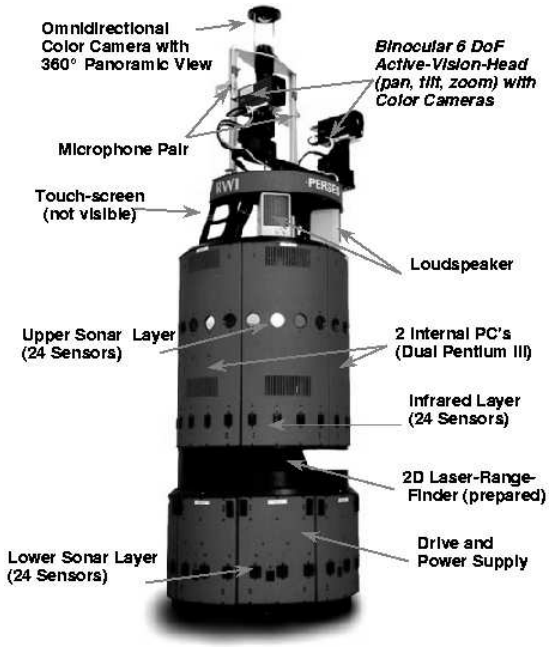

Figure 3: Experimental platform Perses.

2.2 Movement Detection within the Omnidirectional Images

For every mobile service robot, one major problem consists in the robust localization of a potential user in its operation area. Our vision-based user localization 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 (see fig. 4). The implemented method is

Figure 4: Motion-based segmentation of potential users in the image sequence of the omnidirectional camera. (Left) original image. (Right) segmented image, the two regions correspond with two people at different distances to the robot.

similar to that suggested in the *Pfinder* system [19], 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 continuously adjusted. The model simplification led to a lower computational load resulting in a performance speed-up, surprisingly al-

most without any lost 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. 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 circular direction. These heuristics are needed to determine what direction the most attractive one could be. The concerning 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 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 ist detected gives a measure how long the turn to that direction the robot will take. For the case that several people surrond 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 localization of a potential user clapping her hands or shouting a command, we developed a biologically inspired model of binaural sound localization using interaural time differences and spikes as temporal coding principle [12]. This subsystem realizes (i) the detection of the sound direction in the horizontal half-planes by processing the interaural 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 [15].

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 independent of one another. Consequently, a potential user can attract robot's attention via ego-motion or, alternatively, by emitting a sound.

2.5 Person Verification

To evaluate if there really is a person and if she could be willing to interact with the robot, we developed a verification system that integrates different visual cues. This system should highlight that regions most likely cover the upper part of a person. Concentrating on the upper part of a person has thenfollowing reasons: One has less difficulties concerning (partly) occlusions, and the features described below are very person-specific as well as indicate if the person is roughly aligned towards the robot. Execution of person verification is triggered, when the robot was turned by the localization module. Fig. 5 gives an overview over 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

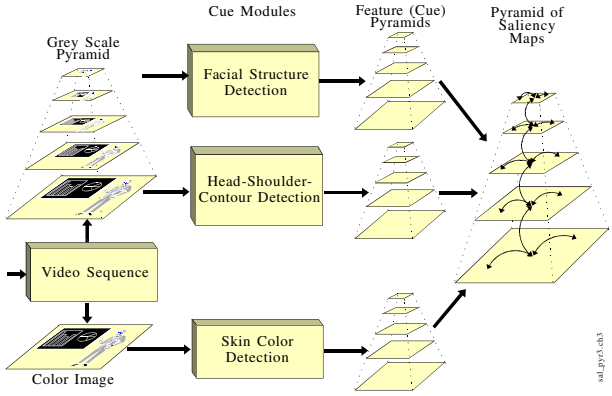

Figure 5: Multiple-cue approach to user verification.

localize people even at different distances from the robot, a multiresolution pyramid (scale space with five fine-to-coarse resolutions) transforms the images into a multiscale representation. Two cue modules sensitive to *facial structure* and *structure of a head-shoulder contour*, respectively, operate at all levels of the grayscale 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 verification system robust and independent of the presence of one certain information source in the images. Hence, we can handle varying environmental circumstances much easier, which, for instance, make the skin color detection

difficult or almost impossible.

For person verification and the subsequent tracking process, both camera systems (omnidirectional as well as frontally aligned stereo system) are utilized. For simplicity reasons, with the exception of movement detection, the examples in this paper contain only images acquired by one of the frontally aligned cameras.

Contour Modelling: The contour which we refer to is that of the upper body of a frontally aligned person. First, we generated a statistically determined average head-shoulder contour by collecting views of different people (see fig. 6). The arrangement itself was learned based on this set of training images. Our simple contour shape prototype model consists of an arrangement of oriented filters realizing a piecewise approximation of the upper shape of a person (head, shoulder). Applying such a filter arrangement in a multi-resolutional manner leads to a robust localization of frontally aligned people even in depth. For computing the

Figure 6: Illustration of the statistically determined contour model by the binary contour shape (top) and the local orientation values along the contour (bottom). Orientation angles are coded by gray values (0^o: black; 90^o: medium gray; 180^o: white).

orientation along the contour, a method proposed in [10] was implemented. Compared to classical orientation-specific filtering with Gabor wavelets [11] or steerable filters [8], this method is faster by orders of magnitude. Orientation filtering provides a tuple 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 (tupels) of such a jet code different dominant orientations, dependent on the applied bandpass filter. For contour detection, 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. Distance measure and jet representation allow a two-step coarse-to-fine search in orientation space. First, a preselection 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. Results for head-shoulder contour detection are shown in fig. 14.

Skin Color Detection: Skin color is a typical feature for person detection and person tracking. Usually, a color space where color and intensity information are uncorrelated is employed. A widely used skin color modeling procedure was suggested in [20] 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 modeled by a bivariate normal distribution (fig. 7).

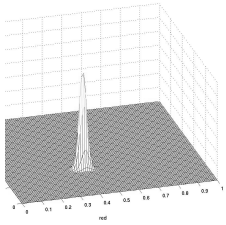

Figure 7: Skin color distribution in r-g color space.

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. 8). 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(x)-height(y) relation is checked according to the expected face region. Subsequently, regions that do not fit the applied criteria can be rejected. Fig. 8 depicts an example for the described skin-color processing regime.

Face Detection: Several approaches to face detection have been described, ranging from using Eigenfaces [18], feature based [21, 5] and neural network based methods [13]. 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 [13], 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.

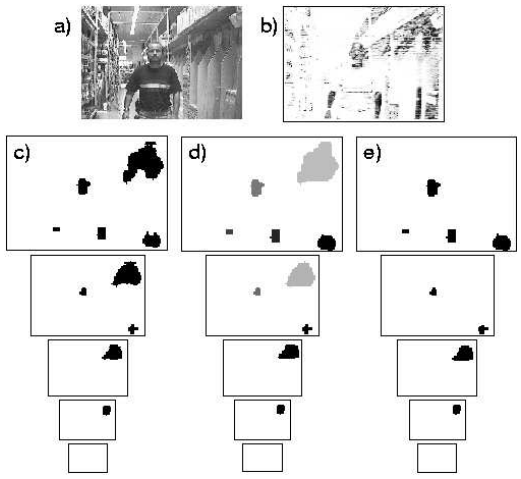

Figure 8: Processing steps for skin color detection: a) original image, b) raw skin color classification, c) smoothing by applying of a median filter for all resolution levels, d) result of the segmentation algorithm, and e) final detection result according to the choosen heuristics for every resolution level.

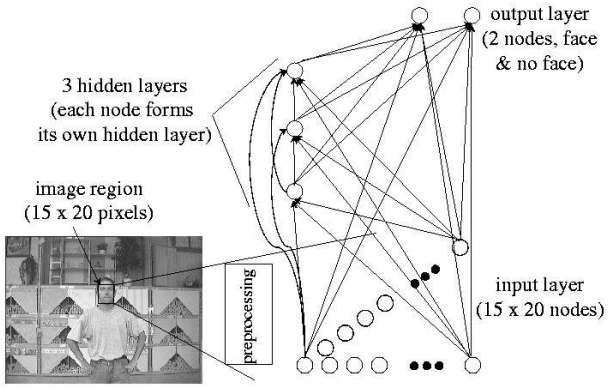

Figure 9: Topology of the CCNW used for face detection.

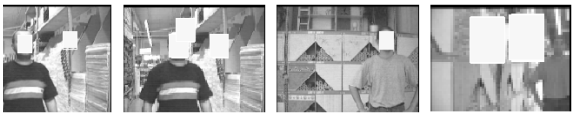

Figure 10: Exemplary results for face detection with the Cascade-Correlation Neural Network, infront of highly cluttered background. From left to right: two images containing correct and false positive detections, an image with only correct detection, and an image where the face detection failed (only false positive detections).

The module for face detection is implemented as a Cascade-Correlation Neural Network (CCNW, [4]). The reason for using that kind of neural network lies in its cabability to produce a network topology that fits optimally with the complexity of the mapping problem. In contrast to the standard Multilayer Perceptron, where the network topology has to be choosen in advance, the CCNW 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, as long as a choosen termination criterion is fulfilled. Fig. 9 depicts the finally obtained topology for the CCNW.

To generate a training pattern set for face images, a public data base provided by AT&T Laboratories Cambridge (http://www.cam-orl.co.uk/facetatabase.html) was utilized. From these images, 15×20 pixel sized regions covering only the face were manually extracted. Initially, the nonface pattern set contains a collection of randomly chosen images, and is extended during bootstrapping. An exemplary result obtained with the CCNW-face detector is shown in fig. 10, further examples can be found in fig. 14. 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.

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 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. 11). 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. Since dynamic neural fields are powerful tools for dynamic selection using simple homogeneous internal interaction rules [1], we adapted them for our purposes. Because we use five fine-to-coarse resolutions in our scale space (see fig. 5), 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 recurrent 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 [3]).

By using a three-dimensional neural field, we are able to consider the local correspondences within as well

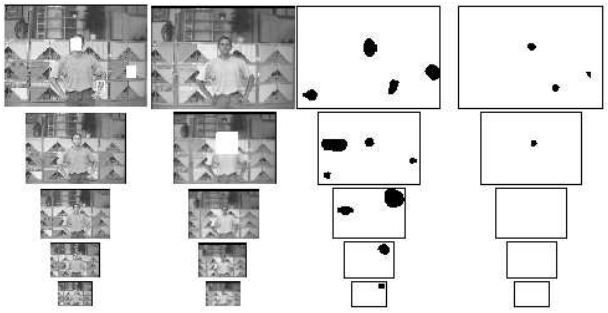

Figure 11: One example for fusing the different cues for person verification. According to the fusion rule, only those locations of the scale space are fed into the final selection process were at least two cues supply a detection result. (From left to right): face detection, head-shoulder contour detection, skin color detection, and fusion result.

as between adjacent resolution levels. This leads to an interesting side effect: because outputs of the different cue detectors offen occur at the same location of adjacent resolution levels, such correspondencies enhance the selection of such locations, resulting in a much more robust verification. The verification examples of fig.14 illustrate this effect.

2.6 Tracking

The goal of person tracking is to keep continuous contact to the current user, and person verification provides the initialization for the subsequent tracking process. Tracking can be done via the omnidirectional camera as well as via the frontally aligned cameras. The tracking procedure is based on the *Condensation* algorithm [9], widely accepted as a powerful and efficient method for tracking arbitrarily shaped probability distributions [7].

The features underlying the tracking process, a combination of head-shoulder contour detection and adaptive color modeling turned out to be appropriate, were derived from the presented person verification procedure.

2.7 Grafical User Interface, Speech Output and Robotic Face

Via the grafical user interface, running on a touchscreen 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 chose 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 verification the robot welcomes this potential user and invites her to interact via the touch-screen.

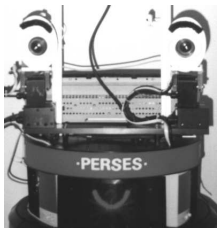

Figure 12: Face of PERSES.

Inspired by the smart face of MINERVA, the robotic tour-guide described in [17], our robot PERSES was equipped with its own face, created by eye-like camera fronts and a mouth made of a controllable diode array (see fig. 12). Hence, the current "emotional" state of the robot can be transmitted in a more natural way.

3. Experimental Results

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

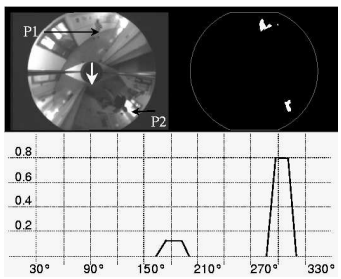

Figure 13: Person localization via motion detection. The white arrow marks the view direction (direction of the frontally aligned cameras, 0^o) of the robot, the angle runs clockwise.

Interaction starts with person localization. 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.

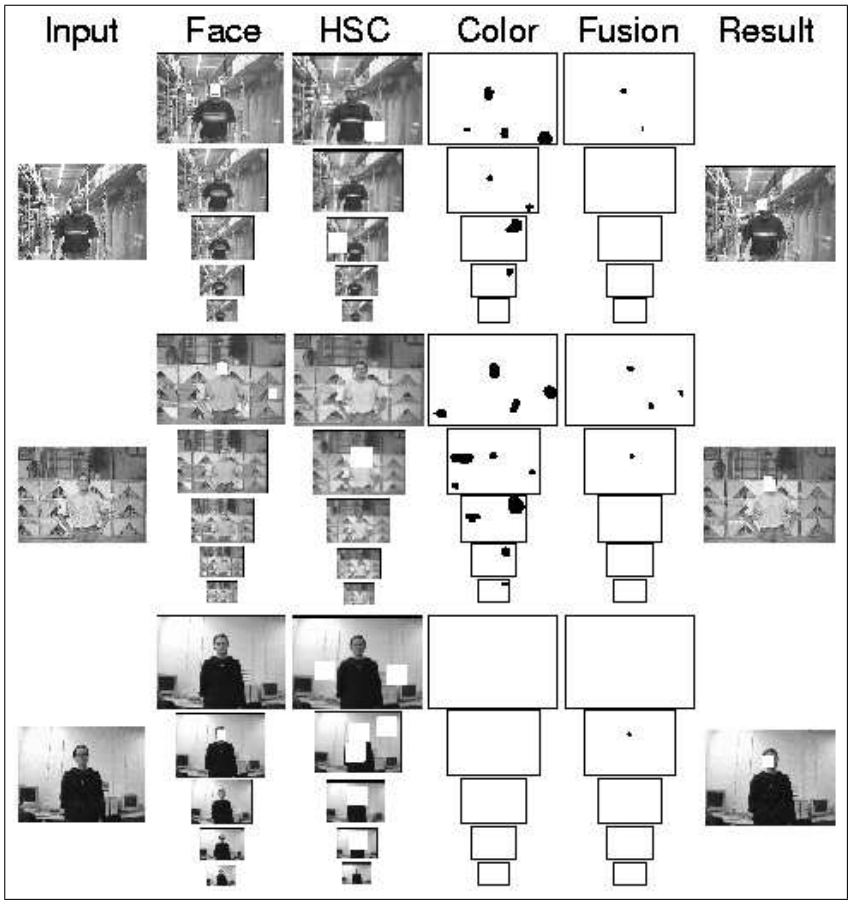

Figure 14: Verification results for different situations: in a hallway of the home store (top), infront of a highly structured wall-paper shelf of the home store (middle), and infront of a relatively uniform background in our lab (bottom). The input image is shown left, the rightmost image contains the verification result. Between them (from left to right) face, head-shoulder contour, and skin color detection are visible. The rightmost scale space column illustrates the different fusion levels after applying the above mentioned fusion rule. During the last year, special emphasis has been made to improve robustness, specificity and efficiency of the verification 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).

Person localization is followed by person verification. Fig. 14 contains a collection of verification results. The verification module provides an output only when cue fusion and final selection (see section 2.5) supply a very strong result. To avoid false verifications is very important, because otherwise the robot would start interaction with uninterested people or even with inanimate items. In case a potential customer has not been verified, the customer can log-in directly via the touch-screen. After successful verification, the robot welcomes the customer by means of a typical speech sequence. Then, the customer can chose the desired item or the interesting market area by means of 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 customer 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. To improve the robustness of the tracking procedure, the person verification module runs in parallel and supplies additional candidate points. Furthermore, the tracking process is supported via a simple sonar-based distance check in the corresponding direction.

Figure 15: Person tracking experiment. The frames indicate those locations where the tracked person is most likely expected.

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, it provides a speech output to ask the user to reduce the distance to the robot. Alternatively, the guidance to a desired market position is interrupted and the robot moves towards its present user to prevent losing contact.

4. Conclusions and Outlook

The paper has described a multimodal scheme for intelligent and natural human-robot interaction. Special emphasis was placed on vision-based methods for user localization, person verification and person tracking. The interaction regime presented here should be understood as work in progress, under-

going 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 [16]. 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Fox, D., 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.

[8] 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.

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

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

[11] 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.

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

[13] 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.

[14] Roy, D. and Pentland, A. Multimodal Adaptive Interfaces. Technical report, M.I.T. Media Lab Perceptual Computing Section, 1997. TR 438.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] Yang, J., Lu, W., and Waibel, A. Skin-Color Modeling and Adaptation. Technical report, Carnegie Mellon University, 1997. CMU-CS-97-146.

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