in: Proc. 2006 IEEE World Congress on Comput. Intell. & Int. Joint Conf. on Neural Networks (IJCNN'06), pp. 8325-8333, Omnipress

Probabilistic Multi-modal People Tracker and Monocular Pointing Pose Estimator for Visual Instruction of Mobile Robot Assistants

Horst-Michael Gross, *Member, IEEE*
Jan Richarz, Steffen Mueller, Andrea Scheidig, Christian Martin
Ilmenau Technical University, Department of Neuroinformatics and Cognitive Robotics, 98684 Ilmenau, Germany

***Abstract*— In this paper, we present two important aspects of our human-robot communication interface which is being developed in the context of our long-term research framework PERSES dealing with the development of highly interactive mobile robotic assistants. First, we introduce a multi-modal people detection and tracking system, a fundamental prerequisite for the observation of a human interaction partner and his non-verbal instructions given by pointing poses, gestures, head pose and eye gaze. Based on this detection and tracking system, we present a hierarchical neural architecture that is capable of estimating a target point at the floor given a pointing pose, thus enabling a user to command his mobile robot to a specific target position in his local surroundings by means of pointing. In this context, we were especially interested in determining whether it is possible to accomplish such a target point estimator using only monocular images of low-cost cameras. Both the tracker and the target point estimator were implemented and experimentally investigated on our mobile robotic assistant HOROS. The achieved recognition results presented finally demonstrate that it is in fact possible to realize a user-independent pointing pose estimation using monocular images only, but further efforts are necessary to improve the robustness of this approach for everyday application.**

I. INTRODUCTION

In recent years, a lot of research has been done to develop mobile robotic assistants that can interact with - and be controlled by - non-instructed users, making them suitable for application in everyday life. To achieve this, it is essential to integrate man-machine-interfaces that are naturally and intuitively to use. In our ongoing long-term research framework PERSES (PERsonal SErvice Systems) we aim to develop such highly interactive mobile robotic assistants for a wide spectrum of demanding everyday life applications, like shopping assistants for supermarkets or home stores [3], [4] or mobile information kiosks for public buildings or areas [8], [9]. From the human-robot interaction (HRI) point of view, such an interactive mobile service robot must be able to autonomously observe its operation area, to detect, localize, and contact potential users, to interact with them continuously, and to adequately offer its specific services considering the current status of the ongoing dialog. Specific service tasks we want to tackle in this research framework are to interactively guide users to desired areas, rooms or people within its operation area *(guidance function)*, or to follow the user as a smart user-oriented mobile assistant that is able to continuously observe the user and to immediately react on his/her instructions *(service companion function)*. To be a really smart companion, a robot should be able to analyze the gender, age, and facial expression of its interaction partner, to interpret his body/head pose and his movement trajectory (see Fig. 1), and to continuously adapt its dialog strategies and presentation modes to that specific user. In this paper, we will only focus on two important aspects of HRI, the robust multi-modal people detection and tracking and the video-based recognition of pointing poses allowing to command a robot to a specific target position in the local surroundings of the user. Besides the methodical background of these techniques, we are presenting results of a series of experiments obtained with our mobile experimental robot HOROS (HOme RObot System).

This work is partially supported by TMWFK-Grant #B509-03007 to H.M. Gross and a HWP-Grant to A. Scheidig

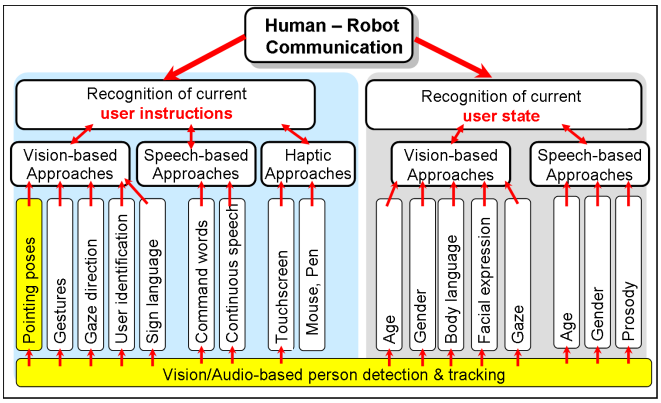

Fig. 1. Systematic overview of those topics in human-robot communication with direct relevance for our long-term research framework PERSES (PERsonal SErvice Systems). The methods presented in this paper can be assigned to the highlighted topics "Person detection" and "Pointing poses".

HOROS' hardware platform is an extended Pioneer II-based robot from ActiveMedia. It integrates an on-board PC (Pentium M, 1.6 GHz) and is equipped with a laser-range-finder and sonar sensors (see Fig. 2). For the purpose of HRI, the robot was equipped with different interaction-oriented modalities. This includes a tablet PC for touch-based interaction, speech recognition and speech generation. HOROS was further extended by a simple robot face which integrates an omnidirectional fisheye camera situated in the center of the head, a camera with a telephoto lens mounted on a tilting socket on the "forehead", and a wide-angle camera in one of the eyes. Because one objective of the PERSES-project is the development of a low-cost prototype of a mobile and interactive robot assistant, we are espe-

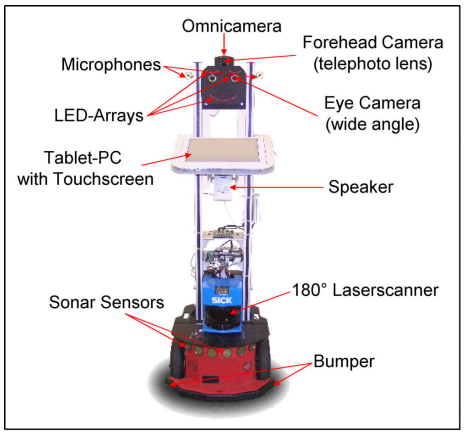

Fig. 2. Equipment of the interaction-oriented robot HOROS.

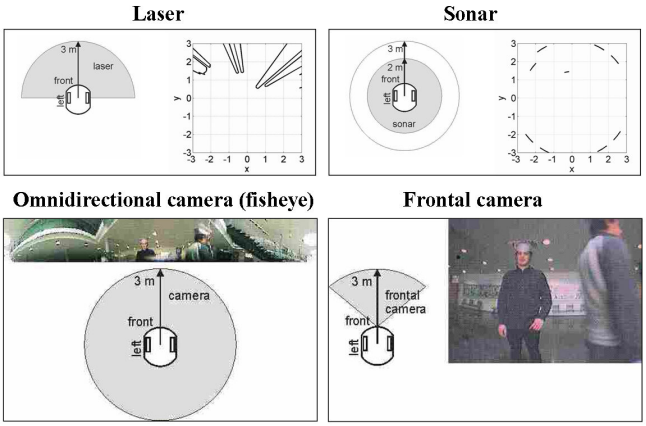

Fig. 3. Exemplary sensory inputs from the laser-range-finder and the sonar (top) and from the fisheye camera and the frontal camera (bottom) for a typical situation, where two people were standing in front of HOROS (see Fig.4, top). In the pictures of the laser and the sonar scan, the robot is located at the 0,0 coordinate straightened ahead. Together with the exemplary inputs the used range of each sensory cue to detect people is depicted. As can be seen, each sensor covers a specific area around the robot. Consequently best tracking results are to be expected if all sensory cues are used concurrently.

cially interested in vision technologies with a good price-performance ratio. Therefore, for the two frontal cameras instead of a stereo-vision system only low-cost cameras were utilized. This forces us to develop powerful and robust recognition algorithms allowing to compensate the deficits of the hardware. In this context, we were interested if it would be possible to robustly estimate a target position at the floor from a pointing pose using only inexpensive hardware and monocular images. A fundamental prerequisite for the recognition of video-based user instructions is, however, a stable detection and tracking of the interaction partner in the local surroundings of the robot. This aspect is described in the following section.

II. MULTIMODAL PEOPLE DETECTION & TRACKING

Typical approaches use visual cues for face detection, a laser-range-finder for detection of moving objects, like legs, or acoustical cues for voice detection. Projects like EMBASSI [2], which aim to detect only the users' faces, usually in front of a stationary system like a PC, typically use visual cues (skin-color-based approaches, sometimes in combination with the detection of edge oriented features). Therefore, these approaches cannot be applied for a mobile robot which has to deal with moving people with faces not always perceivable. Other approaches, e.g. TOURBOT [15] or GRACE [17] trying to perceive the whole person rather than only the face use laser-range-finders to detect people as moving objects. Drawbacks of these approaches occur, for instance, in situations where a person stands near a wall and cannot be distinguished from the background, in scenarios with objects yielding leg-like scans, like table- or chair-legs, or if the laser-range-finder does not cover the whole 360^o.

For real-world scenarios, more promising approaches combine more than one sensory channel, like visual cues and the scan of the laser-range-finder. An example for these approaches is the SIG robot [10], which combines visual and auditory cues. People are detected by a face detection system and tracked by using stereo vision and sound source detection. Further examples are the EXPO-ROBOTS [16], where people are detected as moving objects by a laser-range-finder (resulting from differences from a given static environment map) firstly. After that, these hypotheses are verified by visual cues. Other projects like BIRON [1] detect people by using the laser-range-finder to find leg-profiles and combine these information with visual and auditory cues. The essential drawback of most of these approaches is the sequential integration of the sensory cues. People are detected by laser information only and are subsequently verified by visual or auditory cues. These approaches typically fail if the laser-range-finder yields no information, for instance, in situations when only the face of a person is perceivable because of leg occlusion.

Therefore, we recently developed a new approach for the integration of several sensor modalities and presented a multi-modal, probability-based people detection and tracking system and its application using the different sensory systems of our mobile interaction robot HOROS. This approach can be characterized by the fact that all used sensory cues are concurrently processed and integrated into a robot-centered map using a probabilistic aggregation scheme. The overall computational complexity of our approach scales very well with the number of sensors and modalities. Up to now we utilize the laser-range-finder, the sonar sensors, the omni-directional and the frontal eye-camera of our experimental platform HOROS (see Fig. 2). The laser-range-finder is a very precise but relatively expensive sensor perceiving the frontal 180^o field of HOROS. Because it is mounted on the robot approximately 30 cm above the ground, it can only perceive the legs of people (see Fig. 3, top left). Not least due to this fact and for reasons of cost we'll pass on it in future. Furthermore, HOROS has 16 sonar sensors arranged approximately 20 cm above the ground and covering the complete 360° field of view. As third sensory cue we use the omnidirectional camera with a fisheye lens also yielding a 360^o view around the robot. An example

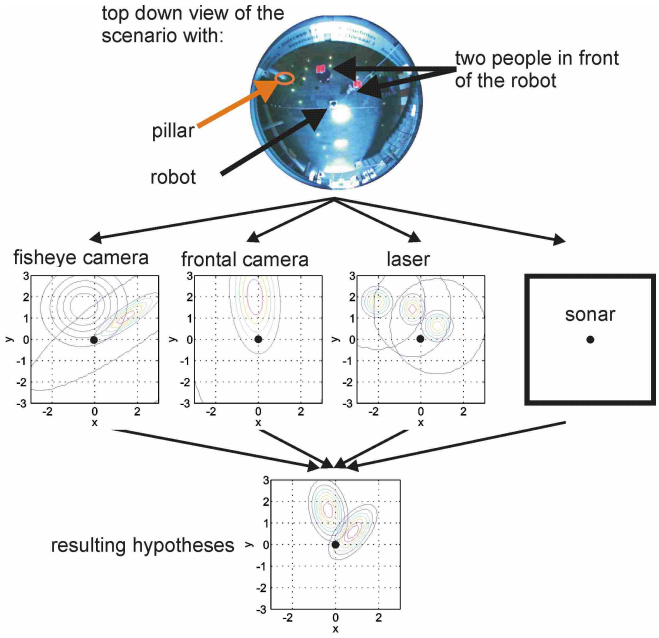

Fig. 4. Aggregation example: (Top) real scene from a bird's eye view - two people are standing in front of the robot. (Middle) current hypotheses generated by fisheye camera, frontal camera, laser-range-finder, and sonar (had to be left out because of technical problems with this sensor system). No sensor on its own can represent the situation correctly. (Bottom) aggregated result from the sensors and the previous timestep. This is a correct and sharpened representation of the current situation.

of an image resulting from this camera is given in Fig. 3 (bottom left). To detect people in the omnidirectional image, a skin-color-based multi-target-tracker [20] is used. This tracking system is based on the condensation algorithm [5] which has been extended to allow the visual tracking of multiple objects at the same time. As fourth sensory cue a frontal camera yielding an approximately $90°$ frontal view is utilized (see Fig. 3, bottom right). To detect people in this image with a head pose oriented to the robot, we use the well-known Adaboost-based face detector of VIOLA and JONES [19]. Similar to the omnidirectional camera, these monocular images only yield hypotheses about the direction of a person but not about his/her distance. Therefore, for these observations we only code the direction of interesting objects and assume constant distances as hypotheses. A more detailed discussion of the advantages and drawbacks of the several sensory modalities is given in [9]. In the nearest future, the results of the sound-source detector utilized to localize a calling person in the local surroundings of the robot and the hypotheses of the head-shoulder detector used to exactly place the Region-of-interest (ROI) for the pointing pose estimation (see Section III) will be integrated into this tracking system. Subsequently, the general idea of the developed approach for probabilistic aggregation of several sensory observations in a robot-centered map is presented.

A. Aggregation and Tracking of Object Hypotheses

For the purpose of tracking, the sensor-specific information about detected human-like objects is converted into Gaussian distributions $\phi(\mu, C)$. The mean μ equals the position of the detection in robot-centered Cartesian coordinates, and the covariance matrix C represents the uncertainty about this position. The form of the covariance matrix is sensor-dependent due to the different sensor characteristics sketched above and described in detail in [9]. All computation is done in the robot-centered x, y space. Examples for the resulting distributions are shown in Fig. 4, middle and bottom. The laser-range-finder yields the most precise data, hence the corresponding covariances are small and the distribution is narrow (see Figure 4, middle). The mean value of the Gaussian depends on the distance of the detected leg-pair and both variances are fixed with approximately 0.4 m. Information from the sonar tends to be very noisy, imprecise and unreliable. Therefore, the variances are large and the impact on the certainty of a hypothesis is lower. In contrast to distance measuring sensors, the cameras can only provide information about the direction of a detection, but not about the distance of a person. Therefore, these Gaussians are modelled with fixed mean distance values - 1.5 m for the fisheye camera and 2 m for the frontal camera. For the variance of the distance, a large value of 1.0 m was selected for the Gaussians of both cameras. The variance in angular direction was also chosen as fixed for the frontal camera with a value of 0.6 m. For the fisheye camera, this variance is directly determined by the angular variance of the particle distribution generated by the skin-color based multi-person-tracker yielding the visual detection hypotheses [20].

Tracking based on probabilistic methods attempts to improve the estimate x_t of the position of a person at time t. These estimates x_t are part of a local map or model M that contains all hypotheses around the robot (Fig. 5). This map is used to aggregate the several hypotheses from the different sensor systems. Therefore, the movements of the robot $\{u_1, ..., u_t\}$ and the observations of humans $\{z_1, ..., z_t\}$ have to be taken into account. In other words, the posterior $p(x_t|u_1, z_1, ..., u_t, z_t)$ is estimated. The whole process is assumed to be Markovian. So, the probability can be computed from the previous state probability $p(x_{t-1})$, the last executed action u_t and the current observation z_t. This way the posterior is simplified to $p(x_t|u_t, z_t)$. After applying the Bayes rule, we get

$$p(x_t|u_t, z_t) \propto p(z_t|x_t)p(x_t|u_t) \quad (1)$$

where $p(x_t|u_t)$ can be updated from $p(x_{t-1}|u_{t-1}, z_{t-1})$ using the motion model of the robot and the assumptions about the typical movements of people. In the map or model, a Gaussian mixture $M = \{\mu_i, C_i, w_i | i \in [1, n]\}$ is used to represent the positions of people, where each Gaussian i is the estimate for one person. $\phi_i(\mu_i, C_i)$ is a Gaussian centered at μ_i with the covariance matrix C_i. The weight w_i $(0 < w_i \leq 1)$ is coding the probability to represent a person by the respective Gaussian.

Next, the current sensor-specific hypotheses z_t have to be integrated. If the map M does not contain any elements at time t, all generated hypotheses from z_t are copied to M. Otherwise data association has to be done to determine

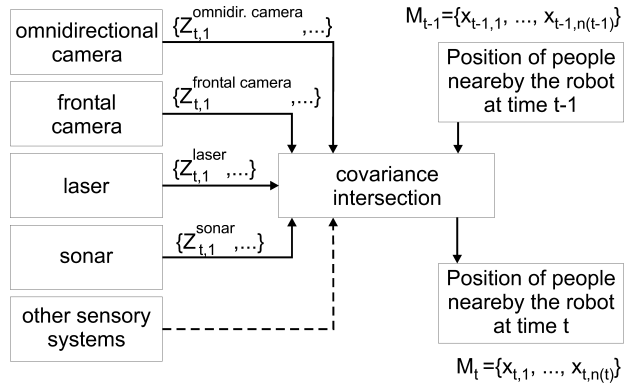

Fig. 5. The architecture of the tracking system: The observations $z_{t,i}^{omni\ camera}$, $z_{t,i}^{frontal\ camera}$, $z_{t,i}^{laser}$ and $z_{t,i}^{sonar}$ of the different sensory cues are combined in a local map M_t that contains a time varying number $n(t)$ of estimates $x_{t,j}$ around the robot using the *Covariance Intersection* rule [7].

which elements from z_t and M refer to the same hypothesis. For that purpose, the Euclidian distance d_e between the respective Gaussians $\phi_i \in z_t$ and $\phi_j \in M$ are used as association criterion. As long as there are distances lower than a threshold, the sensor hypothesis i and the map hypothesis j can be merged. This is done by means of the *Covariance Intersection* rule [7]. By applying this rule, the resulting determinant is minimized by preferring the sharper distribution in the intersection process (see Fig. 4, bottom). With that, an unreliable sensor hypothesis has only little influence on the resulting hypothesis. Sensor readings that do not match with any hypothesis of M are introduced as new hypothesis in M. The weight w_i of a Gaussian is representing the certainty of the respective map hypothesis. The more sensors support this hypothesis, the higher this weight should be. If the weight passes a threshold, the corresponding hypothesis is considered to be a person. In the case of an non-matching hypothesis, the weight is decreased. A person is considered to be lost in the map if t_v seconds passed and no sensor has made a new detection that can be associated with this hypothesis. Mathematical details of this probabilistic aggregation scheme are also given in [9].

B. Experimental Results

To evaluate our multi-modal multi-person tracker we obtained data from an experimental setup, where the robot was standing in a foyer and people moved around it. The environment additionally contained numerous distracting objects, like a pillar and several skin-colored objects. As illustrated in the aggregation example in Fig. 4 no sensor modality alone was able to detect the people and their positions correctly. Only aggregation over several sensor modalities and temporal integration led to the proper result. The whole experimental setup was monitored by a top-down camera mounted 3.5 m above the robot. Because the robot did not move in this experimental setup, we were able to get a reference of the positions of the robot and the persons moving around it (see Fig. 4, top). To determine the performance of the tracker, first the detection rate was evaluated by searching for a tracker hypothesis for each known person position in an image of the top-down camera. Taking into account the noise in the top-down reference, a person was counted as a correct detection if the distance between tracker hypothesis and top-down position was below 50 cm. To get an impression up to what range the tracker is able to find people, the detection rate has been evaluated for different distances of people to the robot. Up to a distance of 2.4 m nearly 80% of all persons in the top-down image have been detected correctly, taking into account that the maximum range of the used sensors to detect people is 3 m. In further experiments the average position error of the trajectories was evaluated. Three typical plots of estimated trajectories with a length above a minimum threshold and the respective top-down trajectories are shown in Fig. 6. The high similarity between ground truth and estimated trajectories with only local displacements is obvious. The position error is typically below 0.5 m, which is sufficient for many subsequent tasks, like the pointing pose estimation, that only require coarse hypotheses where a potential user could be in the local surroundings of the robot. By turning the robot towards the hypothesis with the largest weight (defined, e.g., by the distance to the robot), the potential user can be directly localized in front of the robot allowing the frontal cameras to evaluate if that person could be willing to interact with the robot. As a very simple criterion, we assume that a tracked person may be considered to be a user willing to interact if his upper part of the body is oriented towards the robot. This decision is also taken by means of a Viola & Jones detector - in this case a head-shoulder detector. If this proves to be true, in the next step the robot can try to recognize the user's current state or his given instructions. In the case presented here, we are interested in estimating the target position of a pointing pose triggered by a preceding voice command, like the call "HOROS!", to attract the robot's attention.

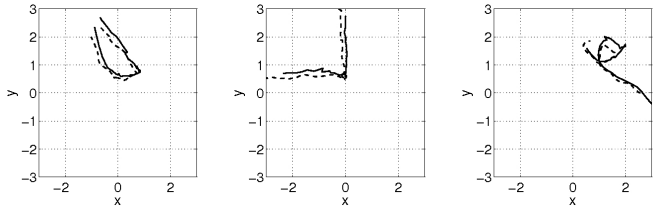

Fig. 6. These pictures show an exemplary comparison of the tracker results (solid line) to the baseline from top-down view (dashed line). The robot was standing in the middle at 0,0 facing upwards.

III. Recognition of Pointing Poses - the Monocular Target Point Estimator

Gestures and poses are a very important aspect of non-verbal inter-human communication. In particular, pointing poses simplify communication by linking speech to objects in the environment in a well-defined way. Therefore, a lot of work has been done in recent years focussing on integrating gesture recognition into man-machine-interfaces.

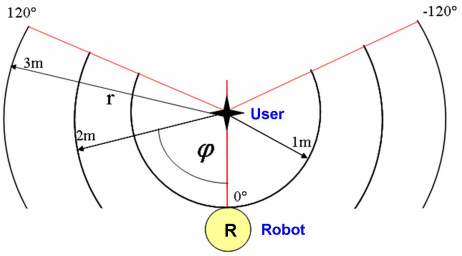

Fig. 7. Configuration used for recording the ground truth training and test data. Here, for reasons of clarity only one of the marked positions in front of the robot to generate pointing poses to predefined target points is shown.

However, most of this work concentrates on distinguishable gestures, creating a "command alphabet" for robot control. Rogalla et al [14], for example, used fourier descriptors of an extracted hand contour and a model database to classify different hand postures. Triesch and v.d. Malsburg [18] detected and classified hand postures by using compound bunch graphs and developed a system that can cope with highly complex backgrounds. Up to now, there are only a few authors who tried to actually estimate a pointing direction out of a deictic gesture. Jojic et al. [6] did so by detecting a person using dense disparity maps and color information. In their approach, a simple Gaussian mixture model is fitted to the person and the pointing direction is determined from the largest principal component of the "arm-blob". Noelker and Ritter [12] used a Local Linear Map (LLM) classifier to detect 2D features in the images of two cameras. A Parametrized Self-Organizing Map (PSOM) then estimates the 3D coordinates of these features, making it possible to calculate a pointing direction. The approach is used to control a Virtual-Reality-System and therefore the working conditions for their system can be very restrictive. Nickel and Stiefelhagen [11] classified dynamic gestures by means of Hidden Markov Models (HMM) and estimated the pointing direction of a pointing gesture by calculating the connecting line between the center of the head and the hand. However, they also used a stereo camera system. With our approach, we were interested to determine whether it is possible to accomplish a pointing position estimator using only monocular images of low-cost cameras as input data. Our goal was to implement that approach on our mobile robot HOROS and make it navigate to specified targets, thus enabling a user to control the robot only by means of pointing. To the best of our knowledge, there are no other low-cost oriented approaches that are comparable to the one presented here.

A. System Overview

1) Pointing Area and Ground Truth Data: We code the target points at the floor as (r, ϕ) coordinates in a user-centered polar coordinate system. This requires a transformation of the target estimate into the robot's coordinate system (by simple trigonometry), but the estimation task becomes independent of the distance between user and robot. Moreover, we limited the valid area for targets to the half space in front of the robot with a value range for r from 1 to 3 m and a value range for ϕ from $-120°$ to $+120°$. The $0°$ direction is defined as user-robot-axis, negative angles are on the user's left side. With respect to a predefined maximum user distance of 2 m, this spans a valid pointing area of approximately 6 by 3 m. Fig. 7 shows the configuration we chose for recording the training data. There are three markers (distance 1, 1.5 and 2 m from the robot) specifying different user positions. Around each marker, three concentric circles with radii of 1, 2 and 3 m are drawn, being marked every $15°$. Positions outside the specified pointing area are not considered. The subjects were asked to point to the markers on the circles in a defined order and an image was recorded each time (see Fig. 8, right). Pointing was performed as a defined pose, with outstretched arm and the user fixating the target point. All captured images are labelled with distance, radius and angle, thus representing the ground truth used for training. This way, we collected a total of 900 images of 10 different interaction partners. During preprocessing, the data were slightly varied to receive nine samples per training image, resulting in a training sample database of 8,100 labelled images.

2) Preprocessing and Feature Extraction: Since the users standing in front of the camera can have different height and distance, an algorithm had to be developed that can calculate a "normalized" region of interest (ROI), resulting in similar subimages for subsequent processing. We use a combination of head-shoulder-detection (based on the Viola & Jones Detector cascade [19] mentioned above), empiric factors, and the distance measurement from the multi-modal person tracker (see Section II) to determine the ROI (Fig. 8). The head-shoulder detector provides a starting point and implicitly includes the user's height into the calculation. For different people, the maximum distances between the center of detection and the tip of the pointing arm in both x and y direction were determined before. These distances were divided by the y-coordinate of the head-shoulder-detection, yielding a factor specifying the size of the ROI. Thus, the size of the extracted image region implicitly depends on the users height. Assuming that the ratio between height and arm length is approximately the same for most humans, this results in an extracted image region that is very similar in

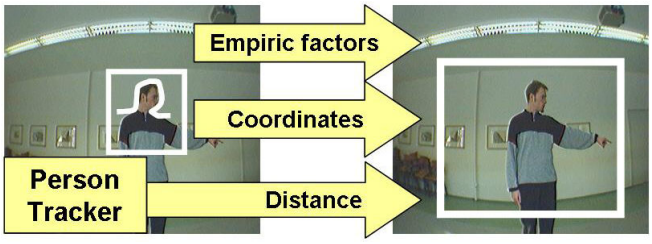

Fig. 8. Example for an image provided by the low-cost eye-webcam. Moreover, this figure sketches how the region of interest (ROI) in the camera image is determined: a combination of empiric factors, head coordinates of the head-shoulder-detector and distance estimation given by the multi-modal tracker (see Section II) is used to achieve a normalized ROI (right).

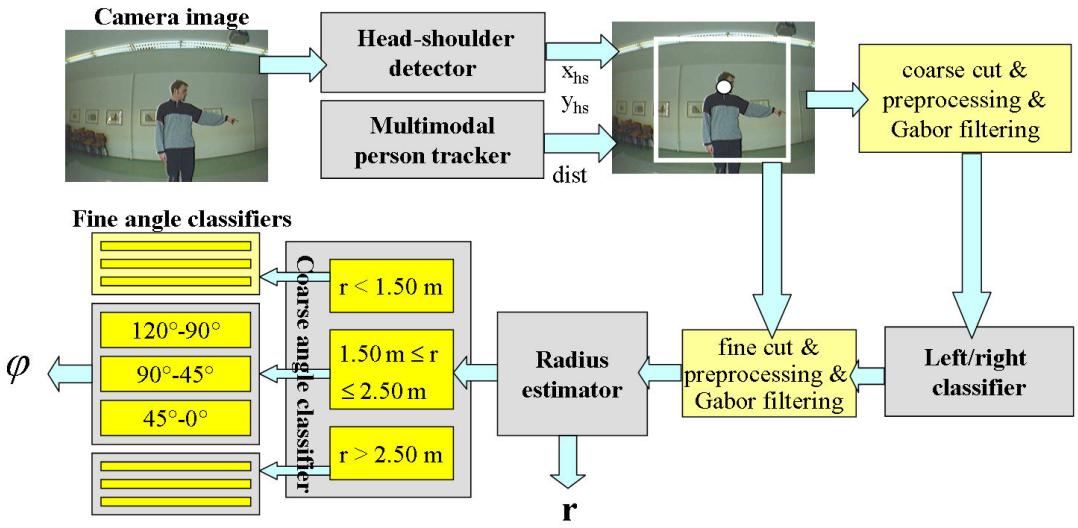

Fig. 10. System overview of the target point estimator cascade. The Gabor-filtered subimage is first fed into a left/right - classifier. The result of this classifier enables it to extract the finer image ROIs shown in Fig. 9, bottom. In the following stage, the final pointing radius r is estimated, and the input is classified into one of three radius classes. For each class, a coarse angle estimator is trained, yielding a classification into one of three angle classes. The last stage yields the final angle estimate ϕ.

Fig. 9. (Top) Captured ROIs extracted with the described normalization algorithm for three instructors with different height (from 1.65 to 2 m) all performing the same pose. (Middle) Extracted ROIs for different distances person-robot ranging from 1-2 m. (Bottom) Examples for sub-images extracted from the ROIs containing both the pointing arm and the head pose. By using these samples as input data for the target point estimator, the head pose is integrated as additional information.

most cases. Finally, the distance estimation from the tracking system allows a simple scaling of the respective ROI. Typical ROIs captured this way are shown in Fig. 9. The found ROI is scaled to 81x81 pixels, and then an illumination correction and histogram equalization is applied. After this, this preprocessed image is Gabor-filtered (4 frequencies with 8 orientations each, absolute values of filter responses) using an equidistant 4x5 grid to extract a pose-describing feature vector as input for the first stage of the pointing estimator. For later stages, the ROI is modified again to create two sub-images, one of them containing the pointing arm, the other one the head (Fig. 9, bottom). By doing this, the head pose of the instructor is directly integrated into the pointing pose estimation as additional information.

3) Architecture of the Classifier Cascade: Experiments showed that it is not possible to tackle the function approximation problem with a single neural network estimating both radius and angle in one step. It also became clear, that while the radius estimation works quite well, it is more difficult to robustly estimate the angle. Therefore we decided to use a cascade of neural classifiers and function approximators (typically three-layered MLPs trained by means of the RPROP learning rule [13]). Fig. 10 gives an overview over the architecture of the developed target point estimator cascade.

After extracting and preprocessing the ROI, a left/right MLP classifier (topology: ((8x4)x(4x5))-40-20-2) first determines whether the person is pointing to the left or to the right. Knowing this, that half of the input image that does not contain the pointing arm can be discarded. This way the ”finer” ROIs containing the head and body-arm regions (see Fig. 9 (bottom)) can be extracted. Each of these two input images is also Gabor-filtered (4 frequencies with 8 orientations, absolute values of filter responses) using an equidistant 5x5 grid resulting in 1,600 input features describing the head and arm pose sub-images. If the person is pointing to the left, the image is simply flipped. This allows to use the same classifier for both directions. In the following cascade stage the value for the pointing radius r is estimated by means of a first MLP function approximator (topology: 1600-30-20-1) with a single output neuron linearly coding the range from 1 to 3 m (output interval: 0 ...1.0). Since the estimation of ϕ is less accurate and prone to errors, this estimation is done later in the cascade, so it can be given as

much supporting and simplifying information as possible. To that purpose, the arm and head ROIs are first classified into one of three coarse radius classes (see Fig. 10, bottom left). For each of these classes, there is a specialized MLP classifier assigning the input to a coarse angle class (topology: 1600-30-20-10-3). Finally, within the respective coarse class, a finer estimation of ϕ is determined by a last MLP function approximator (with slightly different topologies for the 9 sub-classes, typically 1600-20-10-5-1) leading to the final target estimation [r, ϕ]. The cascade contains a total of 14 MLP networks (1x left/right, 1x radius, 3x coarse angles, 9x fine angles), but, due to the hierarchical architecture only four of these classifiers have to be activated during one pass. Calculating all four MLPs takes less than 100 milliseconds on HOROS's onboard computer.

B. Experimental Results

1) Estimation Results of Human Viewers: In order to get a reference value for the recognition performance of the estimator, in particular experiments we determined how accurate a human viewer could estimate the referred target point from a monocular image. Therefore, the images from the training and test data sets were presented to test viewers in random order using the graphical user interface shown in Fig. 11. The valid area for the pointing targets is specified by a circle segment. Subjects were told beforehand that the targets can only lie within this area. The circle segment is skewed perspectively to create a 3D impression and adapted in size according to the distance between the person in the image and the camera. The test person marked a guessed target point by clicking with the mouse pointer on the interface. The found coordinates were then transformed according to the given perspective and the distance of the person, yielding the estimated target coordinates r and ϕ. The estimates were then compared with the known image labels. These comparing experiments were performed with 8 test viewers resulting in 885 target estimates altogether. The achieved estimation accuracy is shown in Fig. 12. On the top, the mean values and standard deviations of the angle estimates are shown over the correct angle. Obviously, perfect estimates would lie on a straight line depicted by the dotted red (gray) line in the image. The mean values of the estimates deviate slightly from this ideal case. It is noticeable that angle estimates between (+/-) 45° and 90° are persistently too large in magnitude. What's more, the standard deviations (depicted by the vertical lines) for these angles are significantly higher. So, it seems to be quite difficult for a human viewer to precisely estimate ϕ from the monocular images in this area. At the bottom of Fig. 12, the average errors for the estimates of r and ϕ over the correct angle are shown. For the radius r, the errors are significantly higher for small angle values compared to large angle values. The errors for ϕ behave inversely, being small for small angle values, then getting bigger with increasing angle value. This trends can be easily explained geometrically. For angles greater than 90°, the errors decrease again. This is due to the fact that pointing to a target behind ones position results in a significant change of the body pose: The shoulder and the face are turned backwards, which is clearly visible in the sub-images. Overall, in 50.1% of all cases, the human viewers estimated ϕ correctly within a tolerance of 10°. For r, 76.3% of all trials were within a tolerance of 50 cm. These results give a hint for valuating the following results of our neural estimator, keeping in mind that the presented data, the distorted monocular images, are very unfamiliar for a human.

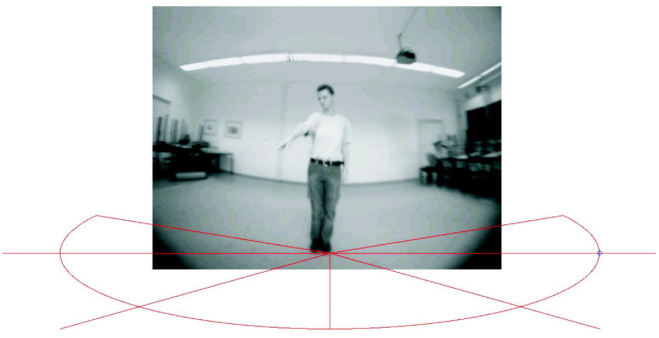

Fig. 11. Graphical user interface for experiments with human viewers. The valid target area is shown by the circle segment.

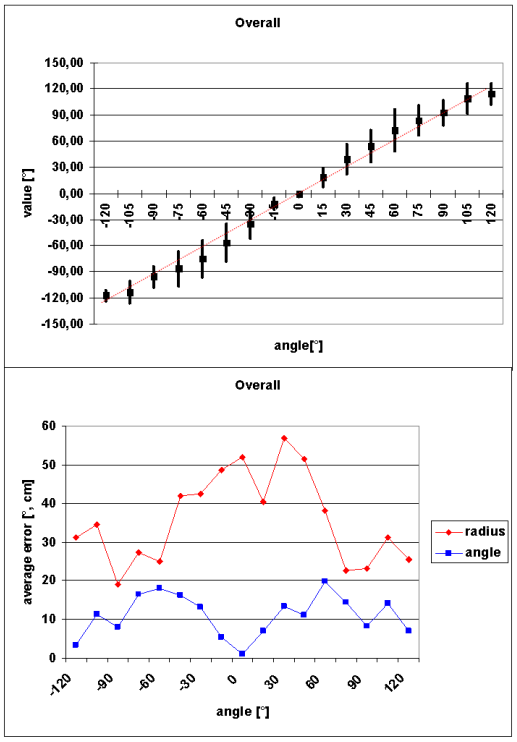

Fig. 12. Estimate results of human viewers. (Top) Mean values and standard deviations of the angle estimate over the correct angle. (Bottom) Average errors of the radius and angle estimates over the correct angle.

2) Results of the Neural Estimator Cascade: In the following experiment, the correct face position was labelled manually in all test images. By means of this step, the

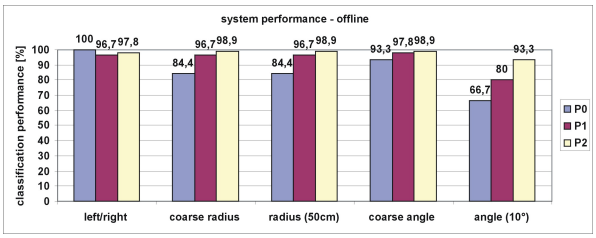

Fig. 13. Classification results of the different stages of the estimator cascade for 3 test subjects. (From left to right:) left/right classifier, radius classifier (radius classes), radius estimate (tolerance 50cm), coarse angle classifier, angle estimate (tolerance 10°)

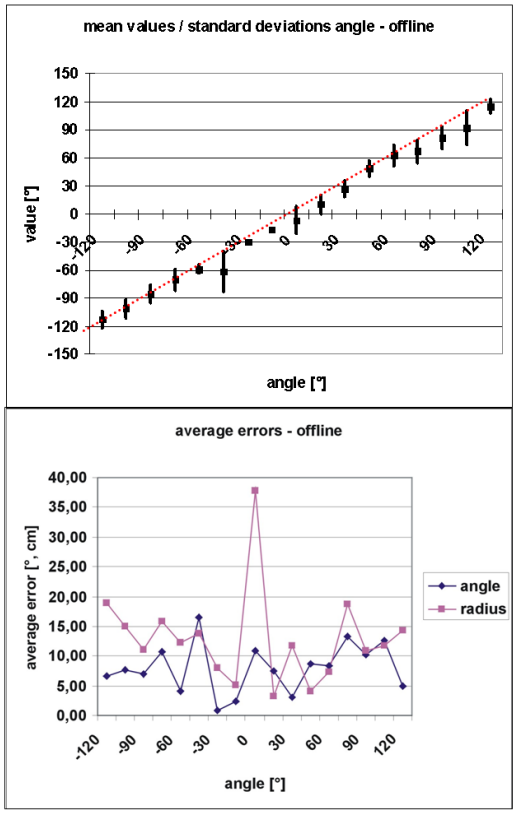

Fig. 14. (Top) Mean values and standard deviations of the angle estimate over the correct angle determined on hand-placed ROIs (off-line estimation). The ideal case is depicted by the dotted line. (Bottom) Average errors for radius and angle estimate over the correct angle.

negative influence of positioning errors possibly generated by the automatic head-shoulder-detection could be completely eliminated. This way, the performance and properties of the developed ROI extraction algorithm and the neural estimator cascade could be analyzed without impairments by deficits of preceding subsystems. Fig.13 shows the classification results of each cascade stage for three test persons. For comparison, person P2 is taken from the training data set. All results mentioned in the following passages refer to the two remaining subjects not included in the training data. The left/right classifier yields classification rates of almost 100% for all subjects. This is especially important since further processing of the input image depends on the results of this stage, and misclassifications will lead to a totally erroneous

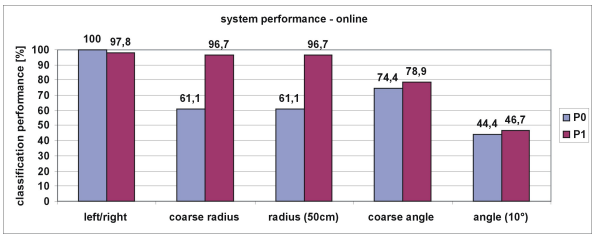

Fig. 15. Online classification results of the different stages of the estimator cascade for two test subjects. In these experiments the head-shoulder-detector was activated for positioning of the ROI.

target estimate. The radius estimator stage shows a good overall performance, with 84.4% and 98.9% of the samples within the 50 cm tolerance and classified into the correct radius class. Compared to this, the angle estimator stages perform poorly: While the performance of the coarse angle classifier stage is very good for all subjects, the fine angle estimate is not so good, with only 66.7/80% of the samples within the 10° tolerance. These results show that the angle estimate is the major problem, limiting the performance and accuracy of the developed pointing direction estimator. For comparison with Fig. 12, the diagram for the mean values and standard deviations of the angle estimate is given in Fig. 14 (top). The results are close to the optimal straight line with small standard deviations for most angles. Fig. 14 (bottom) shows the average errors of the angle and radius estimates. The behavior of the angle estimate is quite similar to that observed in Fig. 12. The radius estimate behaves almost inversely to that observed before, apart from the large errors for 0°. Looking at Fig. 13 again, it can be seen that the neural estimator achieved a classification rate of 66.7% and 80% respectively for the fine angle estimate with a tolerance of 10°, and 84.4% and 98.9% for the radius estimate with a tolerance of 50 cm. This is significantly better than the results achieved by human viewers (50.1 / 76.3%). But of course, the latter are more reliable in the sense that they don't produce outliers and large errors. When interpreting this results, we have to keep in mind that they were achieved off-line with a perfect head detection. Therefore, Fig. 15 demonstrates the performance of the classifier stages for the two test persons when the Viola & Jones detector is activated and used online for head-shoulder-detection. In this case the recognition rate for the coarse radius becomes about 20% lower for person P0 and stays constant for P1, while the fine angle estimates (with a tolerance of 10°) get significantly worse for both persons (only 45%). This clarifies that of all possible error sources, the head-shoulder detection is the most crucial: misplacements of a few pixels from the optimal position may already lead to greater errors in the final target estimate.

To determine the overall online performance and precision of the presented target point estimator while operating on the mobile robot HOROS, a random target pointing experiment was conducted finally: Standing at many different positions within the operation area, the instructor pointed to randomly selected target positions in his local surroundings, and the

robot had to navigate from its current rest position to the estimated target position. From a total of 72 trials, only six (8.3%) were totally erroneous outliers. The remaining trials yielded an average position error of 59 cm. 28 (38.9%) were within 50 cm, 31 (43.1%) within 1 meter, and 7 (9.7%) within 1-2 m from the target point. For a correct interpretation of these results it should be taken into account that in this experiment all possible disturbances and localization errors did superimpose: an imperfect person tracking and head-shoulder detection resulting in non-optimal placed ROIs, an erroneous target point estimation with many different reasons (changing background, badly executed pointing poses, image disturbances, etc.), and insufficiencies in the robot's navigation system resulting, for example, in an imperfect self-localization and motion planning to the given target points.

IV. Summary and Outlook

In the first part of the paper, we presented a multi-modal probability-based approach for detecting and tracking people. It is implemented on our mobile interactive robot HOROS and is working in real-time (7-10 update cycles per second). Because of the sensor fusion and the probabilistic aggregation scheme, its detection and tracking results are significantly improved compared to known single sensor tracking approaches. In our future work, we will extend our multi-modal tracking system with additional cues to further increase robustness and reliability for real-world environments. For example, we are currently integrating the voice command triggering the pointing pose estimation to allow a voice-driven speaker localization, too.

In addition to this multi-modal people tracking approach, we developed a neural classifier cascade for appearance-based estimation of a referred target point at the floor out of a pointing pose. Although we only use monocular image data of relatively poor quality, the system accomplishes a good target point estimation, achieving an accuracy better than that of a human viewer on the same data. The achieved performance rates demonstrate that it is in fact possible to realize a user-independent pointing pose estimation using monocular images only, but further efforts are necessary to improve the robustness of this approach for everyday application.

There are several possible improvements to our system that need to be investigated in the near future: First, the used feature extraction (Gabor filtering using an equidistant grid) seems to be too simple. Several more sophisticated methods for feature extraction and representation are imaginable that may lead to better results. For instance, a foreground extraction routine, e.g. based on active contours or shapes, could be applied, segmenting the pointing person from the background and thus limiting disturbing background influences. Second, further efforts are necessary to improve the accuracy of the head-shoulder detection preceding the target point estimation. Possibly this can be achieved by combination with the active contours allowing to compensate the deficits of a simple input-driven detector. Moreover, so far we only evaluated the performance of our target point estimator on single images of the final pointing pose. An interesting question is whether the dynamic movement of the pointing arm to the final pose contains additional information that could be exploited to enhance the precision of the estimator. In a first experiment, we utilized a Kalman filtering algorithm with a very simple system model. The results suggest that this could indeed improve the estimator performance, especially the accuracy of the angle estimate. However, further investigations are required on this topic.

References

[1] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. Fink, G. Sagerer, "Audiovisual person tracking with a mobile robot", in: Proc. Int. Conf. on Intelligent Autonomous Systems (IAS), IAS Press, 2004, pp. 898–906.

[2] B. Froeba, C. Kueblbeck, "Real-time face detection using edge-orientation matching", in: Proc. Audio- and Video-based Biometric Person Authentication (AVBPA'2001), 2001, pp. 78–83.

[3] H.-M. Gross, H.-J. Boehme, "PERSES - a Vision-based Interactive Mobile Shopping Assistant", in: Proc. 2000 IEEE Intern. Conf. on Systems, Man and Cybernetics (IEEE-SMC 2000), pp. 80–85

[4] H.-M. Gross, A. Koenig, H.-J. Boehme, Chr. Schroeter, "Vision-Based Monte Carlo Self-localization for a Mobile Service Robot Acting as Shopping Assistant in a Home Store", in: Proc. 2002 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS 2002), pp. 256–262

[5] M. Isard, A. Blake, "Condensation - conditional density propagation for visual tracking", Int. Journal on Computer Vision 29 (1998) 5–28.

[6] N. Jojic, B. Brumitt, B. Meyers, S. Harris, T. Huang, "Detection and estimation of pointing gestures in dense disparity maps", in Proc. Int. Conf. on Automatic Face and Gesture Recognition, 2000, pp. 468-475.

[7] S. Julier, J. Uhlmann, "A nondivergent estimation algorithm in the presence of unknown correlations", in: Proc. American Control Conference, Vol. 4, IEEE, 1997, pp. 2369–2373.

[8] C. Martin, H.-J. Boehme, H.-M. Gross, "Conception and realization of a multi-sensory interactive mobile office guide", in: Proc. IEEE Conf. on Systems, Man and Cybernetics (SMC), 2004, pp. 5368–5373.

[9] C. Martin, E. Schaffernicht, A. Scheidig, H.-M. Gross, "Sensor Fusion using a Probabilistic Aggregation Scheme for People Detection and Tracking", in: Proc. of the 2nd European Conference on Mobile Robots (ECMR 2005), pp. 176–181, stampalibri 2005

[10] K. Nakadai, H. Okuno, H. Kitano, "Auditory fovea based speech separation and its application to dialog system", in: Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Systems (IROS), 2002, pp. 1320–1325.

[11] K. Nickel, R. Stiefelhagen, "Real-time recognition of 3D pointing gestures for human-robot-interaction", in Proc. DAGM-Symposium 2003, pp. 557-565.

[12] C. Noelker, H. Ritter, "Illumination independent recognition of deictic arm postures", in Proc. 24th Annual Conf. of the IEEE Industrial Electronics Society 1998, pp. 2006-2011.

[13] M. Riedmiller, H. Braun, "A direct adaptive method for faster back-propagation learning: the RPROP algorithm", in Proc. ICNN-93, San Francisco, 1993, pp. 586-591

[14] O. Rogalla, M. Ehrenmann, R. Zoellner, R. Becher, R. Dillmann, "Using gesture and speech control for commanding a robot assistant", in Proc. 2002 IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN 2002), pp. 454-459.

[15] D. Schulz, W. Burgard, D. Fox, A. Cremers, "Tracking multiple moving objects with a mobile robot", in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2001, pp. 371–377.

[16] R. Siegwart et al., "Robox at expo.02: A large scale installation of personal robots", 42 (2003) 203–222.

[17] R. Simmons et al., Grace: An autonomouse robot for AAAI robot challenge, AAAI Magazine 24 (2) (2003) 51–72.

[18] J. Triesch, Chr. v.d. Malsburg, "Classification of hand postures against complex backgrounds using elastic graph matching", in Image and Vision Computing, volume 20 (2002), pages 937-943.

[19] P. Viola, M. Jones, "Rapid object detection using a boosted cascade of simple features", in Proc. Conference of Computer Vision an Pattern Recognition (CVPR), 2001, vol. 1, pp. 511-518.

[20] T. Wilhelm, H.-J. Boehme, H.-M. Gross, "A multi-modal system for tracking and analyzing faces on a mobile robot", in: Robotics and Autonomous Systems, Vol. 48, 2004, pp. 31–40.