In Proceedings of IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China, May 2011

Towards Joint Attention for a Domestic Service Robot – Person Awareness and Gesture Recognition using Time-of-Flight Cameras

David Droeschel, Jörg Stückler, Dirk Holz, and Sven Behnke

Abstract— **Joint attention between a human user and a robot is essential for effective human-robot interaction. In this work, we propose an approach to person awareness and to the perception of showing and pointing gestures for a domestic service robot. In contrast to previous work, we do not require the person to be at a predefined position, but instead actively approach and orient towards the communication partner. For perceiving showing and pointing gestures and for estimating the pointing direction a Time-of-Flight camera is used. Estimated pointing directions and shown objects are matched to objects in the robot's environment.**

Both the perception of showing and pointing gestures as well as the accurary of estimated pointing directions have been evaluated in a set of different experiments. The results show that both gestures are adequatly perceived by the robot. Furthermore, our system achieves a higher accuracy in estimating the pointing direction than is reported in the literature for a stereo-based system. In addition, the overall system has been successfully tested in two international RoboCup@Home competitions and the 2010 ICRA Mobile Manipulation Challenge.

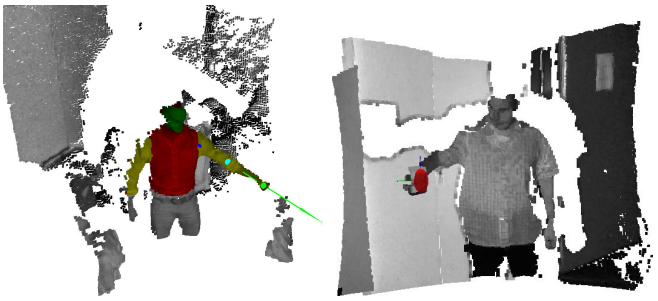
I. Introduction

The requirements for service robots differ vastly from those of industrial robots. Service robots need to work in everyday environments, in close interaction with humans. For effective human-robot interaction, joint attention [1], [2] is essential. Joint attention refers to the ability to selectively attend to an object of mutual interest. Only when both communication partners refer to the same object in their environment, they can exchange about this object.

Showing and especially pointing gestures are a common and intuitive way to draw somebody's attention to a certain object. However, establishing joint attention between humans and robots solely based on gestures is highly asymmetric. While humans can easily interpret robot gestures [3], the perception of human behavior using robot sensors is more difficult. Humans have a large repertoire of social cues, such as gaze direction, pointing gestures, and postural cues, that all indicate to an observer which object is currently under consideration.

In this work, we propose for a domestic service robot an approach to joint attention that combines person awareness with the perception of pointing and showing gestures. In contrast to previous work [4], we do not require the person to be at a predefined position. Using laser-range finders (LRFs) and cameras, our robot detects and keeps track of persons in its surrounding. To this end, we employ visual verification of person hypotheses tracked in the LRF data and active gaze control strategies. For communication, persons are approached and looked at by the robot. In order to draw the robot's attention to a particular object, the person can simply point towards the object's location or show the object to the robot (see Fig. 1).

All authors are with the Autonomous Intelligent Systems Group, University of Bonn, Germany. Email: droeschel@ais.uni-bonn.de



(a) Pointing gesture (b) Showing gesture

Fig. 1. Recognizing pointing and showing gestures. a) The user points to an object in the scene. b) The user shows an object to the robot.

We use a Time-of-Flight camera as the primary sensor for perceiving gestures. For pointing gestures, the pointing direction is estimated and matched with objects in the robot's environment. In the case of showing gestures, the robot tries to extract and visually recognize the shown object. By perceiving these gestures, the robot's attention can be drawn to a certain object.

The remainder of this paper is organized as follows: After a brief review of related work in the respective fields in Section II, we outline our system consisting of methods for detecting and tracking multiple persons (Section III) as well as for perceiving showing gestures (Section IV). In Section V, we focus on the perception of pointing gestures. We evaluate the accuracy of estimated pointing directions and the applicability of the overall system in Section VI.

II. Related Work

Joint attention with robots has been investigated, e.g., by Hafner and Kaplan [5]. They used simple edge-based features to recognize pointing gestures of Aibo dogs in a specific setting. Such a simplified approach is not suitable for the recognition of human gestures. In addition to perceiving gestures, our approach involves detecting, tracking, and approaching communication partners.

A. Person Detection and Tracking

Tracking people with laser-range finders is a well studied topic in mobile robotics (e.g., [6], [7]). Many approaches detect and track legs of people and fuse this information in a multi-hypothesis tracker [8].

The computer vision community developed a variety of methods for tracking multiple persons with camera systems. For statically mounted cameras (e.g., [9], [10]), background subtraction can be applied to improve the robustness of tracking. When the camera moves (as in [11], [12], [13]), subtracting background is no longer possible. Instead, robust person detectors are required that provide stable information for tracking.

B. Recognizing Pointing and Showing Gestures

Gesture recognition has been investigated by many research groups. A recent survey has been compiled by Mitra and Acharya [4]. Most existing approaches are based on video sequences (e.g. [14], [15]), since color cameras provide images at high frame rates. These approaches are, however, sensitive to lighting conditions. We use a Time-of-Flight (ToF) camera for gesture recognition. This active sensor measures depth independent of the lighting. ToF cameras have been used for recognizing hand gestures [16], [17] and human pose estimation [18].

Pointing for grasping on a table has been described by McGuire *et al.* [19]. They use skin color-based segmentation to localize human forearms. Martin *et al.* [20] use background subtraction to improve the estimation of the pointing direction. Their approach starts from face detection and determines two regions of interest, where Gabor filter responses are analyzed. Sumioka *et al.* [21] used motion cues to establish joint attention. In the approach proposed by Nickel *et al.* [14], skin color information is combined with stereo-depth in order to track 3D skin color clusters. To be independent of lighting conditions, the authors initialize the skin color using pixels of detected faces.

Related to our approach of showing objects to the robot is the work of Goerik *et al.* [22]. Their approach is based on a stereo camera system and an initial object segmentation using depth information.

Common to all the above approaches is that they require the person to be at a predefined position, in the field-of-view of the camera. In contrast, our robot actively approaches persons and directs the time-of-flight camera towards them.

Nickel *et al.* [14] use a statically mounted stereo-camera system for perceiving pointing gestures. They apply a color-based detection of hands and head and cluster the found regions based on depth information. Using Hidden Markov Models (HMMs) trained on different phases of sample pointing gestures, they estimate two types of pointing directions – the head-hand line and the 3D forearm direction.

III. Continuous People Awareness

For human-robot interaction, a key prerequisite for a robot is awareness of the whereabouts of people in its surrounding. We combine complementary information from laser range finders (LRFs) and vision to continuously detect and keep track of people (cf. Fig. 2). In LRF scans, the measurable features of persons like the shape of legs are not very distinctive, such that parts of the environment may cause false detections. However, LRFs can be used to detect person candidates, to localize them, and to keep track of them at high rates. In camera images, we can verify that a track belongs to a person by detecting more distinctive human features like faces and upper bodies on the track.

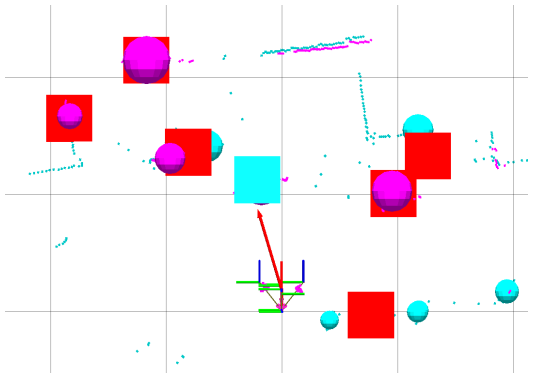


Fig. 2. Persons are detected as legs (cyan spheres) and torsos (magenta spheres) in two laser range scans (cyan and magenta dots). The detections are fused in a multi-hypothesis tracker (red and cyan boxes). Faces are detected with a camera mounted on a pan-tilt unit. We validate tracks as persons (cyan box) when they are closest to the robot and match the line-of-sight towards the face (red arrow). We also determine the face height by projecting the track position onto the face direction.

A. Detection and Tracking of Multiple Persons

Our domestic service robot is equipped with two LRFs. One LRF is mounted shortly above the ground at a height of 24 cm and detects legs of people. We additionally detect torsos of people with a second LRF at a height of approx. 80 cm.

In a multi-hypothesis tracker, we fuse both kinds of detections to tracks. Position and velocity of each track are estimated by Kalman filters (KFs). In the KF prediction step, we use odometry information to compensate for the motion of the robot. After data association, the tracks are corrected with the observations of legs and torsos. We use the Hungarian method [23] to associate each torso detection in a scan uniquely with existing hypotheses. In contrast, as both legs of a person may be detected in a scan, we allow multiple leg detections to be assigned to a hypothesis. Only unassociated torso detections are used to initialize new hypotheses. A new hypothesis is considered a person candidate until it could be verified as a person through vision. Spurious tracks with low detection rates are removed.

B. Person Verification

In order to verify tracks as persons, we extract human features from camera images. One module detects frontal and profile views of faces using the Viola and Jones [24] algorithm. For a detected face, we determine a best matching track that lies closest to the line-of-sight towards the face. We check the scale of the face by requiring that the projection of the track into the image lies horizontally within a multiple of the face detection bounding box. Additionally, we compare the width of the face bounding box with the projected width of a standard sized head at the track's location by imposing upper and lower bounds on the widths' ratio.

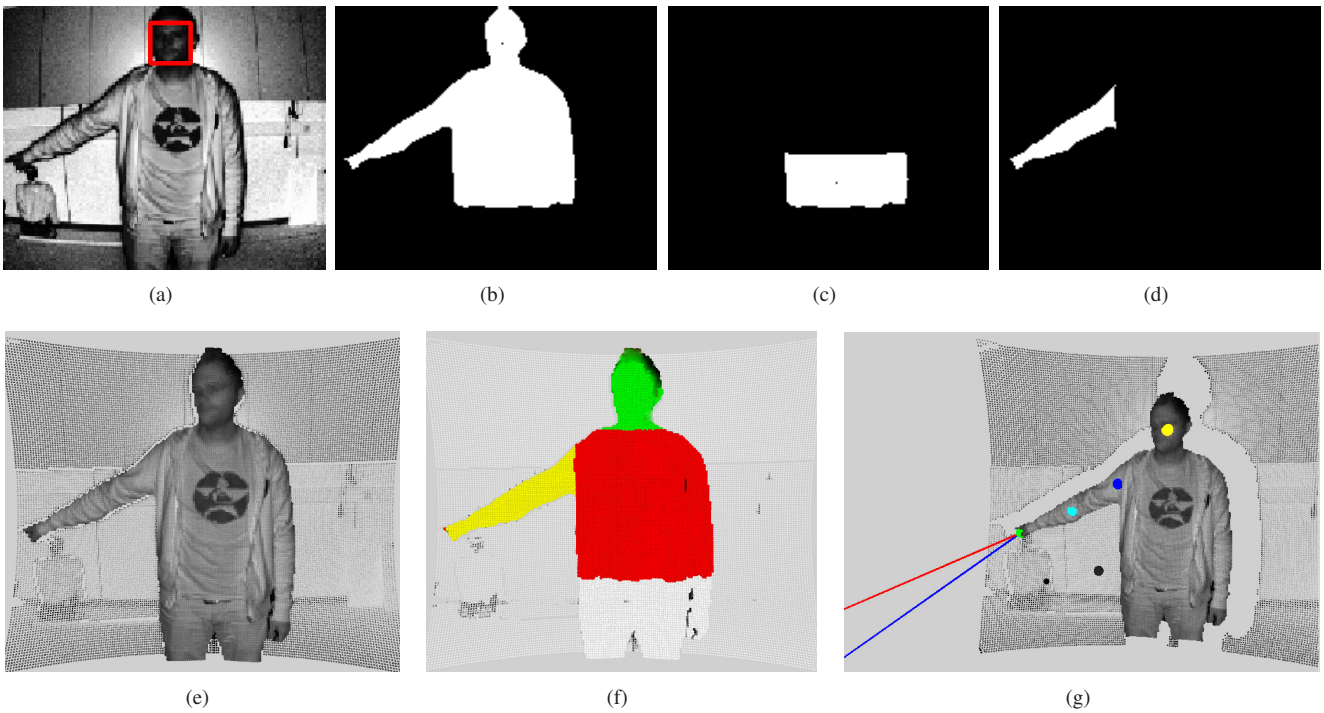


Fig. 3. Results of the segmentation steps. a) The amplitude image with a detected face, depicted by the red bounding box. b) Body segment. c) Abdomen segment. d) Arm segments. e) Unprocessed point cloud with intensity values coded in levels of gray. f) Head (green), torso (red) segment and arm segment (yellow). g) The determined pointing vectors: eye-hand pointing vector (blue arrow) between the face centroid (yellow sphere) and the hand position (green sphere) and the elbow-hand pointing vector (red arrow) between the elbow position (cyan sphere) and hand position (green sphere).

In a second module, we detect upper bodies with a method based on Histograms of Oriented Gradients [25]. As this method is computationally demanding, we project the track into the image and extract an adequately sized image patch at this image location. To capture upper bodies in a large variety of human poses, we determine the patch size by projecting a rectangle around the track's position into the image. Horizontally, we choose the rectangle to be orthogonal to the line-of-sight towards the track and to cover reasonable upper body widths. In the vertical, the rectangle spans a wide range (1 m in our setting) of upper body heights.

C. Gaze Control for Person Verification

Since the LRFs measure in a larger field-of-view (FoV) than the cameras, it is not possible to verify all tracks as persons in a single view. To enhance the FoV of the camera effectively towards the FoV of the LRFs, we implemented an active gaze strategy that utilizes the pan-tilt neck and the yaw joint in the torso of our robot.

Among all tracks, the gaze strategy selects the one with the highest importance. We measure importance of a track as a linear combination of the following criteria:

- Low angular distance of a track from the current view direction in the horizontal plane.
- Proximity to the close-range communication region in front of the robot.
- Distance to a specific number of person verifications. This component is only used if the track has been verified as a person beforehand.
- At least one view on each track.
- Discount track, if it has been viewed multiple times without being verified as person.

IV. Recognition of Showing Gestures

For perceiving showing gestures and recognizing the object that is shown, we use both a regular color camera and a ToF camera. Processing the acquired sensory information is composed of the following steps.

1) Range-Image Processing to Detect Showing Gestures: We first segment the depth image of the ToF camera according to the measured distances. The segments are then used to form three-dimensional clusters in the point cloud.

For all clusters, we determine oriented bounding boxes and neglect those falling below or above a minimum and a maximum size. Furthermore, since the robot has approached the person beforehand, we can assume that clusters belonging to the person or the object are close to the robot. That is, we can neglect clusters being farther away than, e.g., 2 m. As a side note, it is to remark that the object in hand normally occludes parts of the holding hand and arm; thus leading to clusters that are not much larger than the object itself.

As a potential candidate, we select the closest remaining cluster, as shown in Fig. 1(b). The selected object cluster is then tracked using a similar method as in Sec. III-A.

2) Recognition of the Shown Object: For recognizing the object in hand, the object's position and size is projected into the color camera image. For object recognition, we use color histograms and SURF features as described in [26].

3) Gaze Control for Gesture Recognition: One limitation of ToF cameras is their small apex angle. Therefore, we actively control the orientation of the head to keep the relevant parts, i.e., the person's face and hand within the field-of-view.

V. Pointing Gesture Recognition

The perception of pointing gestures is based on amplitude images as well as three-dimensional point clouds of a ToF camera. This allows to perceive the 3D direction in which the person is pointing. We determine the pointing direction in three steps – detecting the person's head, segmenting the person's body into parts, and localizing the person's elbow and hand.

1) Head Detection: In the amplitude image of the ToF camera, we detect frontal and profile views of faces using the Viola and Jones [24] algorithm. Fig. 3(a) shows an amplitude image in which a user faces the camera and performs a pointing gesture. We seek to determine the centroid of the head and approximate this point with the centroid of the points on the head within the face bounding box. When a face is detected, we first determine the centroid of the 3D points within the face bounding box. Since the 2D face bounding box may contain background, we remove outliers from the head cluster by rejecting points with a large distance to the centroid. From the remaining points, we redetermine the head centroid.

The detection performance of the Viola and Jones algorithm is not perfect. Its detection rate, for example, decreases with distance from the frontal or profile view. This occurs frequently during gesture recognition, since people tend to look into the direction they are pointing. We resolve this issue by tracking the head cluster in the 3D point cloud once it has been found through face detection.

2) Body Segmentation: Once the head is detected, we segment the person's body from the background. For this purpose, we apply 3D region growing using the centroid of the head as a seeding point. To accelerate computation, we approximate a point's neighborhood by a 2D pixel neighborhood of the camera's image array. ToF cameras measure a smooth transition where should be depth jump-edges at object boundaries [27]. In order to avoid the merging of unconnected regions, jump-edge filtering is an essential prior to region growing. We terminate region growing if a point exceeds the maximal extensions of a human upper body. We approximate the maximal extensions by a bounding box that extends $100\,\mathrm{cm}$ from the head downwards and $100\,\mathrm{cm}$ in each horizontal direction.

In order to reliably segment the arms from the remainder of the torso, we determine the diameter of the abdomen. We assume that the center of the abdomen is located $50\,\mathrm{cm}$ below the head. Furthermore, if the arms perform a pointing gesture, they are not connected with the abdomen in the point cloud. In this case, we can consider those points of the person's body as belonging to the abdomen that lie below the upper chest, i.e. at least $40\,\mathrm{cm}$ below the head. To obtain the arm segments, we first exclude all points in the body segment that lie within the horizontal projection of the abdomen. Then we grow regions on the remaining points to find the individual arms. Fig. 3 illustrates the main steps of the segmentation procedure as binary images and as 3D point clouds, respectively.

3) Hand and Elbow Localization: To find the arm and elbow locations a cost value for every point in the arm segment is calculated. The cost of a specific point corresponds to the traveled distance from the head during the region growing process. As a result, the finger tip will always have the maximal cost assigned independent of the arm posture. Thus, the hand location is approximated by the centroid of the points with the maximum cost in the arm cluster. The elbow can be found by exploiting the anatomical property that forearm and upperarm have similar length. Hence, the elbow is given by the point with median distance to the head. The shoulder is simply the point from the arm cluster with minimal distance to the head. Fig. 3(g) shows determined locations of hand, elbow, and shoulder in an exemplary situation.

4) Gesture Detection: We segment the pointing gesture in three phases, the preparation phase which is an initial movement before the main gesture, the hold phase which characterizes the gesture, and the retraction phase in which the hand moves back to a resting position. We train Hidden Markov Models (HMMs) for the individual phases. Since gesture phases appear in a given order, the HMMs for the specific phases are composed in a topology similar to [15].

As input to the HMMs we use expressive features extracted in the previous step. The input feature vector f is defined as $f = (r, \phi, v)$, where r is the distance from the head to the hand, ϕ the angle between the arm and the perpendicular body axis and v the velocity of the hand.

5) Determining the Pointing Direction: After detecting a pointing gesture we determine the intended pointing target by calculating the pointing vector that corresponds to the pointing direction of the user. Similar to [14], the *eye-hand* vector and the *elbow-hand* vector are calculated. The *eye-hand* vector corresponds to the line between the estimated face centroid and the hand position. In contrast, the *elbow-hand* vector is the line between the elbow and the hand position which corresponds to the extended forearm (c.f. Fig. 3(g)). The pointing direction is mapped to a specific pointing target by finding the target with the minimum distance to the pointing vector.

VI. Experiments

A. Pointing Direction Evaluation

In order to evaluate the accuracy of the pointing gesture recognition, we conducted experiments in an indoor scenario. 16 test persons have been asked to perform 24 pointing gestures to 20 different pointing targets. The pointing targets have been distributed in the scene at different height levels $0\,\mathrm{m}$, $0.8\,\mathrm{m}$, $1.5\,\mathrm{m}$ and $2\,\mathrm{m}$, with at least $0.5\,\mathrm{m}$ distance. Fig. 4 shows the setup. The participants were instructed to perform separate, natural pointing gestures to a sequence of pointing targets, including some of the targets twice. The

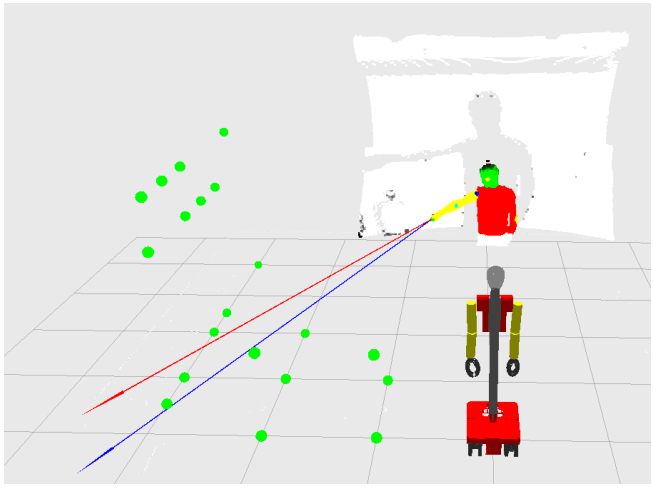


Fig. 4. The experiment setup as perceived by the robot. The test person is in front of the robot at 2 m distance and points to the pointing targets (green spheres). The two pointing directions are depicted by the arrows, Eye-Hand (blue), elbow-hand (red).

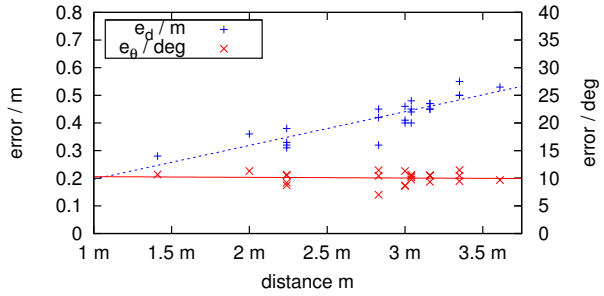


Fig. 5. The average distance (e_d) and angular errors (e_θ) for the eye-hand vector by distance to the pointing target.

order and selection of the pointing targets was randomly chosen, ensuring that the same pointing targets were not in succession. The pointing targets were announced to the test persons one by one right before they performed the pointing gesture, to avoid a prepossession in the pointing direction.

The position of the pointing targets was manually measured. For every pointing gesture we calculated the shortest distance between the pointing vector and the target position e_d and the corresponding angular deviation e_θ for the two pointing vectors.

The overall average error of all pointing gestures and all test persons is shown in Table I. It can be seen that the candidates seem to point in the direction of the eye-hand line and that the eye-hand vector seems to be the better approximation for the pointing direction. An average error of, respectively, 0.43 m and 10.1° is achieved.

TABLE I

POINT GESTURES: AVERAGE DISTANCE AND ANGULAR ERROR

	Avg. e_d/m	σ_d/m	Avg. $e_\theta/$ deg	$\sigma_\theta/$ deg
Eye-Hand	0.43	0.19	10.1	4.38
Elbow-Hand	0.53	0.28	12.58	6.91

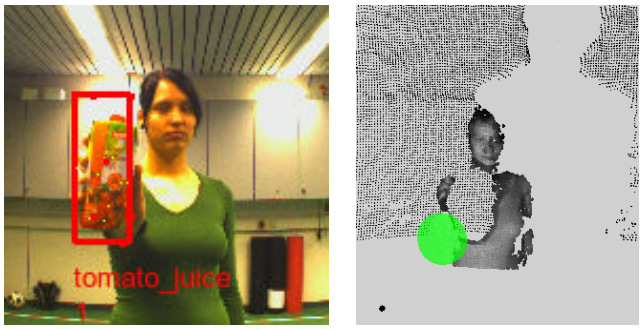


Fig. 6. Perceiving showing gestures and recognizing objects: the object that is shown to the robot is correctly detected (red rectangle/green sphere) and recognized as being a tomato juice.

The results indicate that while the position error e_d increases with the distance to the pointing target, the angular error e_θ seems to be constant, as illustrated in Fig. 5 for the eye-hand vector.

Compared to the approach by Nickel *et al.* [14] that uses a stereo camera system, we achieve a higher accuracy in the pointing target estimation.

B. Pointing Gesture Recognition Evaluation

In order to evaluate the person-independent recognition performance of our system, we split the data from the 16 test persons into a training data set, consisting of 192 pointing gestures from 8 test persons and a test data set, consisting of 192 pointing gestures from the remaining 8 test persons. We train the Hidden Markov Models on the training data set by manually labeling the hold phase of each pointing gesture. From the 192 pointing gestures we identify 187 gestures correctly. For 5 pointing gestures, the hold phase is not detected correctly. In one case, the hold phase was too short. The remaining 4 false detections are caused by an incorrect body segmentation.

C. Showing Gesture Evaluation

In order to evaluate the perception of showing gestures, we conducted an experiment where 5 test persons have been asked to perform 5 showing gestures by choosing an object and showing it to the robot. For all 25 showing gestures, the object was correctly identified by the robot. A typical result can be seen in Fig. 6.

D. RoboCup

Besides the quantitative evaluation of the system, we evaluated our approach with our domestic service robot that competes in the RoboCup@Home league. We successfully applied continuous people awareness and gesture recognition, e.g. during RoboCup German Open 2010 (Magdeburg), RoboCup 2010 (Singapore), and the ICRA 2010 Mobile Manipulation Challenge (Anchorage).

For example in the Who-is-Who test of the RoboCup German Open 2010, the robot detected and identified all 5 persons that either sat or stood in an apartment-like environment. In the finals, the robot successfully approached

and identified a guest. The guest could select something to eat using a pointing gesture to a specific shelf in the environment. The robot detected the pointing gesture, fetched the desired object from the shelf and delivered it to the guest. Also, the guest ordered a new drink by showing his empty drink to the robot. Videos from these experiments are available at [28].

VII. Conclusion

We propose an approach to person awareness and to the perception of pointing and showing gestures for a domestic service robot. Complementary information from laser range finders and vision is used to continuously detect and keep track of people. Once the approaches a person, it perceives two kinds of gestures – pointing and showing gestures, which are recognized using a time-of-flight camera.

Pointing gestures are interpreted by detecting the person's head, segmenting the person's body into parts, and localizing the person's elbow and hand. The pointing direction is matched with locations in the robot's environment. Showing gestures are interpreted by clustering objects in the depth image and neglecting invalid clusters. The robot recognizes the showed object with its color camera. In both cases, the robot fetches the referenced object to the user.

The accuracy of the estimated pointing gestures has been evaluated in an experiment with 16 participants pointing at 24 different objects in a laboratory environment. The results show that our system achieves a higher accuracy in estimating the pointing direction than is reported in the literature for a stereo-based system. Furthermore, we found that the candidates seem to point in the direction of the eye-hand line and that the eye-hand vector seems to be the better approximation for the pointing direction. An average error of $0.43\,\mathrm{m}$ and $10.1°$, respectively is achieved.

Besides that, the pointing gesture recognition system has been evaluated on the same data. From our test data set of 192 pointing gestures we correctly identify 187 gestures, i.e., a detection rate of 0.97%.

In an experiment including 5 participants and 25 showing gestures, we have evaluated the perception of showing gestures. The system could identify all shown objects correctly.

In addition, the overall system has been successfully tested in two international RoboCup@Home competitions and the 2010 ICRA Mobile Manipulation Challenge.

Up to now, our system only supports two gestures, pointing and showing. It is a matter of future work to extend the system to more gesture categories and to dynamic gestures.

Acknowledgment

This work has been supported partially by grant BE 2556/2-3 of German Research Foundation (DFG).

References

[1] L. Kopp and P. Gärdenfors. *Attention as a Minimal Criterion of Intentionality in Robots*, volume 89 of *Cognitive Studies*. Lund University.

[2] F. Kaplan and V. Hafner. The challenges of joint attention. *Interaction Studies*, 7(2):135–169, 2006.

[3] F. Faber, M. Bennewitz, C. Eppner, A. Görög, C. Gonsior, D. Joho, M. Schreiber, and S. Behnke. The humanoid museum tour guide Robotinho. In *Proc. of IEEE Humanoids*, 2009.

[4] S. Mitra and T. Acharya. Gesture recognition: A survey. *IEEE Trans. on Systems, Man and Cybernetis - Part C*, 37(3):311–324, 2007.

[5] V. V. Hafner and F. Kaplan. *Learning to interpret pointing gestures: experiments with four-legged autonomous robots*, volume 3575 of *LNCS*. Springer, 2005.

[6] D. Schulz, W. Burgard, D. Fox, and A.B. Cremers. People tracking with a mobile robot using sample-based joint probabilistic data association filters. *International Journal of Robotics Research*, 22, 2003.

[7] K. Arras, S. Grzonka, M. Luber, and W. Burgard. Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities. In *Proc. ICRA*, 2008.

[8] D.B. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24(6), 1979.

[9] J. Berclaz, F. Fleuret, and P. Fua. Robust People Tracking with Global Trajectory Optimization. In *Proc. of CVPR*, pages 744–750, 2006.

[10] O. Lanz. Approximate bayesian multibody tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(9):1436–1449, 2006.

[11] D.M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *Int. Journal of Computer Vision*, 73(1):41–59, 2007.

[12] A. Ess, B. Leibe, K. Schindler, and L. van Gool. A mobile vision system for robust multi-person tracking. In *Proc. of CVPR*, 2008.

[13] K. Okuma, A. Taleghani, N. de Freitas, J.J. Little, and D.G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *Springer, LNCS 3021*, 2004.

[14] K. Nickel and R. Stiefelhagen. Visual recognition of pointing gestures for humanrobot interaction. *Image and Vision Computing*, 25(12):1875–1884, 2007.

[15] T. Axenbeck, M. Bennewitz, S. Behnke, and W. Burgard. Recognizing complex, parameterized gestures from monocular image sequences. In *Proc. of IEEE Humanoids*, 2008.

[16] E. Kollorz, J. Hornegger, and A. Barke. Gesture recognition with a time-of-flight camera. In *Proc. of Dynamic 3D Imaging DAGM Workshop*, 2007.

[17] P. Breuer, C. Eckes, and S. Müller. Hand gesture recognition with a novel ir time-of-flight range camera a pilot study. In *Proc. of Mirage 2007, Computer Vision / Computer Graphics Collaboration Techniques and Applications*, 2007.

[18] M. Haker, M. Böhme, T. Martinetz, and E. Barth. Self-organizing maps for pose estimation with a time-of-flight camera. In *Proc. of the DAGM 2009 Workshop on Dynamic 3D Imaging*, 2009.

[19] P. McGuire, J. Fritsch, J. J. Steil, F. Rthling, G. A. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter. Multi-modal human-machine communication for instructing robot grasping tasks. In *Proc. of IROS*, 2002.

[20] C. Martin, F.-F. Steege, and H.-M. Gross. Estimation of pointing poses for visual instructing mobile robots under real world conditions. In *Proceedings of 3rd European Conference on Mobile Robots (ECMR), Freiburg*, 2007.

[21] H. Sumioka, K. Hosoda, Y. Yoshikawa, and M. Asada. Acquisition of joint attention through natural interaction utilizing motion cues. *Advanced Robotics*, 21(9), 2007.

[22] C. Goerick, H. Wersing, I. Mikhailova, and M. Dunn. Peripersonal space and object recognition for humanoids. In *Proc. of IEEE Humanoids*, 2005.

[23] H.W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1):83–97, 1955.

[24] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of CVPR*, 2001.

[25] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Prof. of CVPR*, volume 2, pages 886–893, 2005.

[26] J. Stückler and S. Behnke. Integrating Indoor Mobility, Object Manipulation, and Intuitive Interaction for Domestic Service Tasks. In *Proc. of IEEE Humanoids*, 2009.

[27] S. May, D. Droeschel, D. Holz, S. Fuchs, E. Malis, A. Nüchter, and J. Hertzberg. Three-dimensional mapping with time-of-flight cameras. *Journal of Field Robotics, Special Issue on Three-Dimensional Mapping, Part 2*, 26(11-12):934–965, 2009.

[28] Institut für Informatik VI, Rheinische Friedrich-Wilhelms-Universität Bonn. NimbRo@Home. http://www.nimbro.net/@Home/.