In Proceedings of 6th ACM/IEEE International Conference on Human-Robot Int

Learning to Interpret Pointing Gestures with a Time-of-Flight Camera

David Droeschel
droeschel@ais.uni-bonn.de

Jörg Stückler
stueckler@ais.uni-bonn.de

Sven Behnke
behnke@cs.uni-bonn.de

Autonomous Intelligent Systems Group, Computer Science Institute VI
University of Bonn, Bonn, Germany

ABSTRACT

Pointing gestures are a common and intuitive way to draw somebody's attention to a certain object. While humans can easily interpret robot gestures, the perception of human behavior using robot sensors is more difficult.

In this work, we propose a method for perceiving pointing gestures using a Time-of-Flight (ToF) camera. To determine the intended pointing target, frequently the line between a person's eyes and hand is assumed to be the pointing direction. However, since people tend to keep the line-of-sight free while they are pointing, this simple approximation is inadequate. Moreover, depending on the distance and angle to the pointing target, the line between shoulder and hand or elbow and hand may yield better interpretations of the pointing direction. In order to achieve a better estimate, we extract a set of body features from depth and amplitude images of a ToF camera and train a model of pointing directions using Gaussian Process Regression.

We evaluate the accuracy of the estimated pointing direction in a quantitative study. The results show that our learned model achieves far better accuracy than simple criteria like head-hand, shoulder-hand, or elbow-hand line.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics—*Operator Interfaces, Sensors*; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Range data*

General Terms

Design, Human Factors, Measurement

Keywords

Human-Robot Interaction, Gesture Recognition

.

Figure 1: Application scenario from ICRA 2010 Mobile Manipulation Challenge. Our robot approaches the user and he selects a drink by pointing to it.

1. INTRODUCTION

Non-verbal communication is a key component of human-human interaction. Humans use non-verbal cues to accentuate, complement, or substitute spoken language. In the context of human-robot interaction, however, many approaches focus on speech recognition and well-designed dialogues, although the interpretation of non-verbal cues such as body pose, facial expressions, and gestures may either help to disambiguate spoken information or further complement communication [5, 10, 7].

An important non-verbal cue in human communication is pointing. Pointing gestures are a common and intuitive way to draw somebody's attention to a certain object. While robot gestures can be designed in a way that makes them easily understandable by humans [4], the perception and analysis of human behavior using robot sensors is more challenging.

We investigate natural human-robot interaction in the context of domestic service tasks. In our system (cf. Fig. 1, [16]), we complement speech recognition with the interpre-

tation of gestures such as pointing to objects, showing of objects, or stopping the robot. Using a Time-of-Flight (ToF) camera, we recognize pointing gestures and interpret them in order to infer the intended pointing target.

From the depth images of the ToF camera, we extract body features such as the position of the head, the elbow, and the hand. Frequently, the line between a person's eyes and hand is assumed to be the pointing direction. However, since people tend to keep the line-of-sight free while they are pointing, this simple approximation is inadequate. Moreover, depending on the distance and angle to the pointing target, the line between shoulder and hand or elbow and hand may yield better estimates of the pointing direction. Instead of interpreting gestures with such simple approximations, we propose to learn a model of pointing directions from the observation of humans. For this purpose, we use Gaussian Process Regression (GPR, [3]) as a non-parametric function approximator.

We evaluate our approach in an experiment with 16 participants pointing at 24 different objects. The results show that our learned model achieves far better accuracy than simple criteria like the head-hand, shoulder-hand, or elbow-hand lines. We compare our system to state-of-the-art approaches, e. g., using stereo vision, and demonstrate superior performance.

The remainder of this paper is organized as follows: After a review of related work in Section 2, we briefly introduce our service robot system, in which we integrated pointing gestures as a natural cue in human-robot interaction. We detail our approach for recognizing and interpreting pointing gestures in Sections 4 and 5. Finally, we evaluate the accuracy of the estimated pointing directions in Section 6.

2. RELATED WORK

Gesture recognition has been investigated by many research groups. A recent survey has been compiled by Mitra and Acharya [14]. Most existing approaches are based on video sequences (e. g., [7, 15, 1]). These approaches are, however, sensitive to lighting conditions. In contrast, we utilize a Time-of-Flight (ToF) camera which actively illuminates the scene and measures depth independent of the lighting. ToF cameras have already been used to recognize hand gestures [9, 2] and for human pose estimation [6, 8].

Loper et al. [10] recognize two gestures for commanding a robot to halt or to enter a room. They use a ToF camera similar to the sensor in our approach. The supported gestures do not include any further parameters like a pointing direction that have to be estimated.

Pointing to objects on a table in close range has been described by McGuire et al. [13]. They use a multi-layer perceptron classifier to localize hand and finger tips in stereo images and estimate the pointing direction from the finger direction. Martin et al. [11] train neural networks on Gabor filter responses. Their approach starts from face detection and determines two regions of interest, where they extract filter responses after background subtraction. Sumioka et al. [18] used motion cues to establish joint attention.

Huber et al. [7] use a stereo vision system which is actively controlled by a behavior-based gaze controller. They track body features in proximity spaces and determine the pointing direction from the shoulder-hand line. Their system detects pointing gestures by the relative angle between forearm and upper arm. In their experimental setup, two different persons pointed to eight marked positions on the floor. The authors report a mean accuracy of 0.41 m with a standard derivation of 0.17 m.

In the approach proposed by Nickel et al. [15], skin color information is combined with stereo-depth for tracking 3D skin color clusters. In order to be independent of lighting conditions, the authors initialize the skin color using pixels of detected faces. Nickel et al. 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.

The use of multiple modalities to complement and disambiguate individual communication channels has been investigated, e. g., by Fransen et al. [5]. They integrate information from visually perceived pointing gestures, audio, and spoken language for a robot that performs an object retrieval task. The robot has to disambiguate the speaker and the desired object in the conversation. However, no quantitative analysis of their approach to pointing gesture recognition is given.

In our approach, we use depth from a ToF camera and learn the correct interpretation of the pointing direction from human observation.

3. SYSTEM OVERVIEW

We investigate intuitive human-robot interaction with our service robots Dynamaid [16] and Cosero. We designed the robots with an anthropomorphic upper body scheme that supports natural interaction with human users. The robots can synthesize and recognize speech as well as gestures.

We demonstrated our system successfully at the ICRA 2010 Mobile Manipulation Challenge and at RoboCup 2010 in the @Home league, where we came in second. In an exemplary scenario, the robot searches for persons, approaches them, and offers them an object retrieval service, e. g., to fetch beverages. The person can either tell the robot what and where to search or simply point to the object.

For such a task, the robot has to be aware of the persons in its surroundings. We combine complementary information from laser scanners and vision to continuously detect and keep track of people [17]. A laser scanner in a height of 24 cm detects legs, while a laser scanner in the lower torso detects trunks of people. In a multi-hypothesis tracker, we fuse both kinds of detections to tracks. With a camera on the head of the robot, we can verify that a track belongs to a person by detecting more distinctive human features like faces and upper bodies on the track. Since the laser scanners 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 laser scanners, we implement an active gaze strategy that utilizes the pan-tilt neck and the yaw joint in the torso of our robot.

Once a new person is found, the robot approaches the person and offers her/him to fetch a beverage of choice. The robot waits for information about the desired object. We developed a method to recognize pointing gestures and to interpret its intended pointing target with a ToF camera. During a gesture, our robot adjusts its gaze to keep the head and the hand of the user visible in the sensor image.

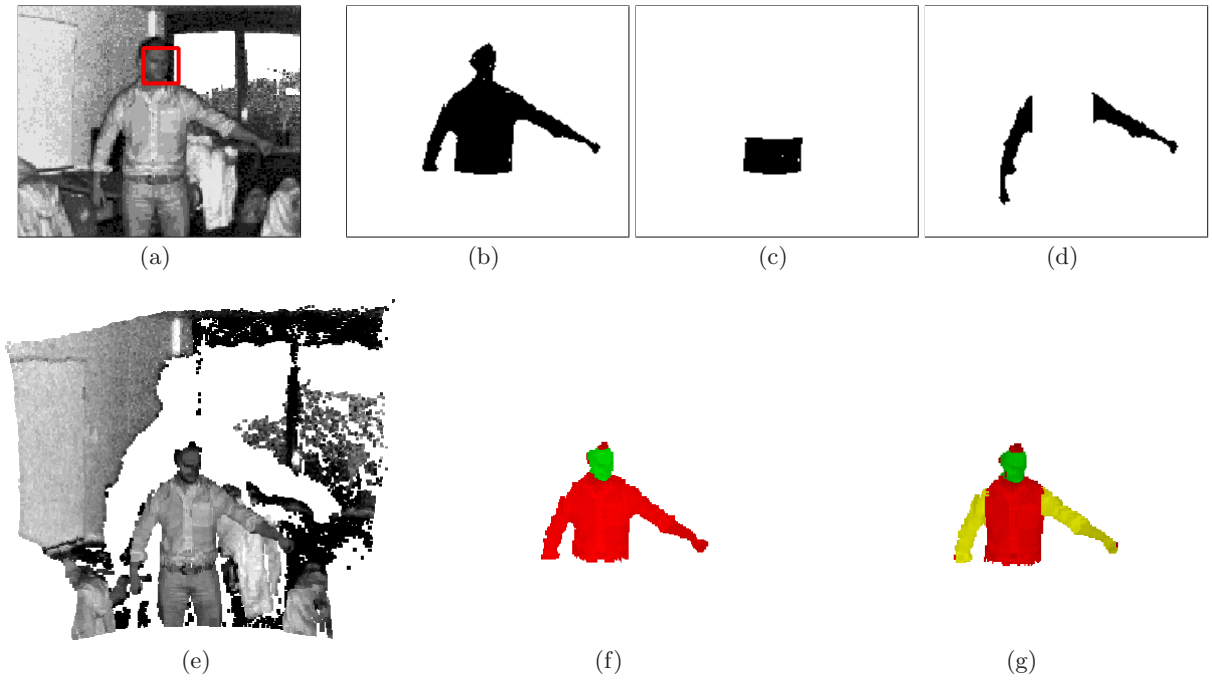

Figure 2: Results of body part segmentation. a) Amplitude image with detected face (red box). b) Body segment. c) Abdomen segment. d) Arm segments. e) Unprocessed point cloud with intensity values coded in levels of gray. f) Head (green) and torso (red) segment. g) Arm segments (yellow).

4. POINTING GESTURE RECOGNITION

Our approach to 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, shoulder and hand.

4.1 Time-of-Flight Camera

Time-of-Flight (ToF) cameras are compact, solid-state sensors, which provide depth and reflectance images at high frame rates. They employ an array of light emitting diodes that illuminate the environment with modulated near-infrared light. The reflected light is received by a CCD/CMOS chip for every pixel in parallel. Depth information is gained by measuring the phase shift between the emitted and the reflected light. The Mesa SR4000, that we are using, also provides the amplitude image that corresponds to the intensity of the reflected light and has similar characteristics to a gray-scale image of a standard visible-light camera.

4.2 Head Detection

In the amplitude image of the ToF camera, we detect frontal and profile views of faces using the Viola and Jones [19] algorithm. Fig. 2 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 as follows.

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. The distance threshold for rejecting points needs to be chosen appropriately to take the sensor's accuracy in distance measurements into account. For our setup, we use a threshold $T_r = 8\,\text{cm}$. From the remaining points, we redetermine the head centroid.

The detection performance of the Viola and Jones algorithm is not perfect. E. g., its detection rate decreases with distance from the frontal or the 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 established through face detection. When a face cannot be found, we initialize the head cluster with the head centroid of the previous frame and keep the resulting head cluster if it is similar in diameter and height.

4.3 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 utilize the 2D pixel neighborhood of the camera's image array. Because ToF cameras measure a smooth transition along depth jump-edges at object boundaries [12], jump-edge filtering is essential prior to region growing in order to avoid the merging of unconnected regions. We terminate region growing if a point exceeds the maximal extensions of a human upper body, described by a bounding box that extends 100 cm from the head downwards and 100 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 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 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. 2 illustrates the main steps of the segmentation procedure.

4.4 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 result, our method assigns a cost value to the finger tip that is close to the maximum cost, 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 upper arm have similar length. Hence, the elbow is given by the point on the arm with median distance to the head. The shoulder is simply the point from the arm cluster with minimal distance to the head. Fig. 3 shows determined locations of hand, elbow, and shoulder in an exemplary situation.

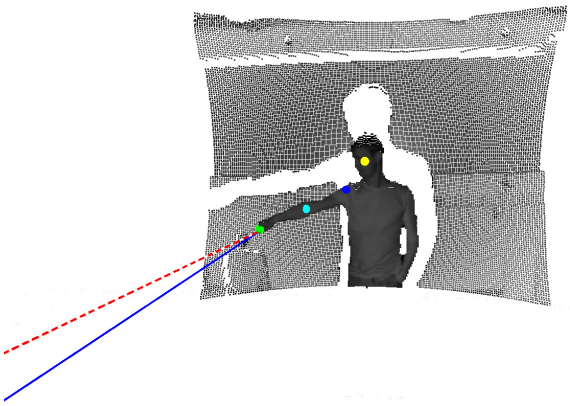

Figure 3: Determined pointing directions: head-hand pointing direction (blue line) between the face centroid (yellow sphere) and the hand position (green sphere) and the elbow-hand pointing direction (red dashed line) between the elbow position (cyan sphere) and hand position (green sphere).

4.5 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 [1].

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 is the angle between the arm and the vertical body axis and v is the velocity of the hand.

5. POINTING DIRECTION ESTIMATION

After a pointing gesture has been detected, we seek to interpret its intended pointing target. The pointing direction strongly depends on the distance and angle to the target. Especially for distant targets, the line through eyes and hand may be used to approximate the line towards the target. However, since people tend to keep the line-of-sight to the target free while they are pointing, this simple approximation is not accurate. Also, the line through shoulder and hand or elbow and hand could provide better approximations for specific target distances and angles.

Instead of interpreting gestures with such simple approximations, we propose to learn a model of pointing directions directly from the observation of humans.

5.1 Gaussian Process Regression

We apply Gaussian Process Regression (GPR, [3]) to train a function approximator that maps extracted body features x to a pointing direction y. The basic assumption underlying Gaussian Processes (GPs) is that for any finite set of points $X = \{x_i\}_{i=1}^N$ the function values $f(X)$ are jointly normally distributed, i. e.,

$$f(X) \sim \mathcal{N}(0, K),$$

where the elements of the covariance matrix K are determined by the kernel function $K_{nm} = k(x_n, x_m)$.

In GPR, observations y_i at points x_i are drawn from the noisy process

$$y_i = f(x_i) + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma_0^2).$$

GPR allows to predict Gaussian estimates for any points x_*, based on training examples $D := \{(x_i, y_i)\}_{i=1}^N$:

$$\mu(x_*) = K_*^T C^{-1} y, \tag{1}$$

$$\sigma^2(x_*) = K_{**} - K_*^T C^{-1} K_*, \tag{2}$$

where $C = K + \sigma_0^2 I$ and $y := (y_1, \ldots, y_N)^T$. The matrices K_{**} and K_* contain the covariances between the query points x_*, and between x_* and the training points X, respectively.

We model similarity in a local context of the input space by means of the Radial Basis kernel function

$$k(x, x') = \theta \; \exp\left(-\frac{1}{2}(x - x')^T \Sigma^{-1} (x - x')\right) \tag{3}$$

with $\Sigma = \text{diag}\left(\sigma_1^2, \ldots, \sigma_M^2\right)$, where $M := \dim(x)$ and θ is the vertical length scale. In regions that are far away from training examples, large predicted variance indicates high uncertainty in the estimate.

In order to perform GPR, we collect training examples of pointing gestures to various pointing locations from several test persons. For a new measurement $\hat{x}$ of body features during a pointing gesture, we apply Eq. 1 to obtain a pointing direction estimate $\hat{y}$ and an associated variance.

5.2 Low-dimensional Gesture Representation

The efficiency of learning algorithms crucially depends on the dimensionality of the parameter space. In order to represent the arm posture in a low-dimensional feature space that

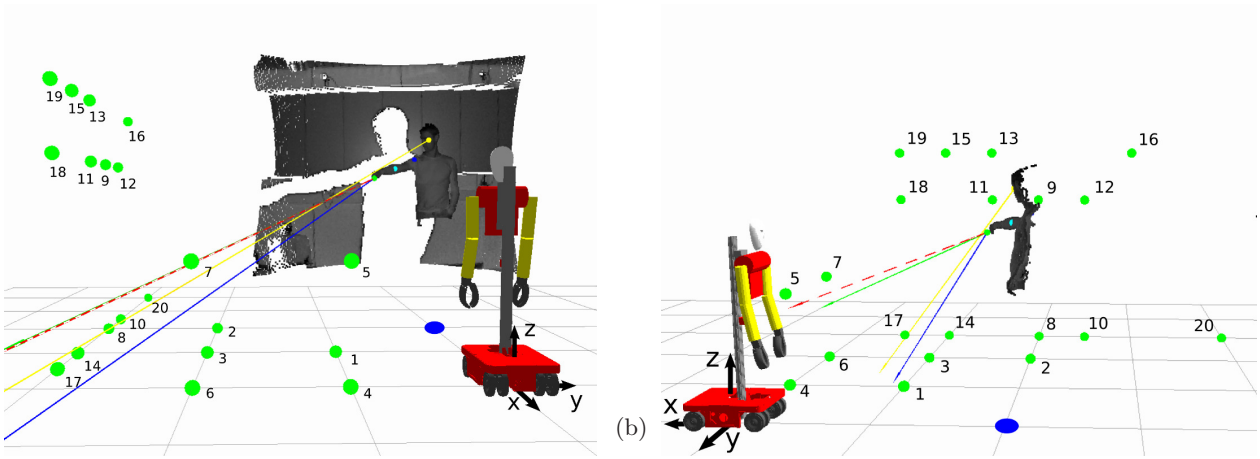

Figure 4: Illustration of experiment setup from (a) frontal and (b) side view. The test person stands in front of the robot at 2 m distance (blue dot) and points to the pointing targets (green spheres). Estimated pointing directions are depicted by arrows: head-hand (blue), elbow-hand (dashed red), shoulder-hand (green), and GPR (yellow). The robot measures angles between body features in its base coordinate frame (black).

is independent of a person's size and arm length, we transform the 3D locations of the hand, elbow, and shoulder into an angular representation.

For this purpose, we measure angles in an egocentric coordinate system from the perspective of robot. In this coordinate system, the x-direction and y-direction point behind and to the right of the robot in the horizontal plane, respectively. The z-axis corresponds to the vertical upward direction. The angular representation consists of yaw and pitch angles of the directions from head to shoulder, from head to elbow, and from head to hand. Hence, the features x in our GPR learning framework are six-dimensional. We model the pointing direction y as yaw and pitch angle relative to the person's head.

6. EXPERIMENTS

In order to evaluate the accuracy of the pointing direction, we conducted experiments in an indoor scenario. We asked 16 test persons with average European body heights and proportions to perform 24 pointing gestures to 20 different pointing targets for our robot Dynamaid [16]. The pointing targets have been distributed in the scene at different height levels 0 m, 0.8 m, 1.5 m and 2 m, with at least 0.5 m distance to each other (cf. Table 1 and Fig. 4). The participants were asked to stand in a distance of 2 m in front of the robot and align their upper body towards it. Fig. 4 illustrates the location of targets, robot, and test subjects in our setup. The participants were instructed to perform separate, natural pointing gestures to a sequence of pointing targets, including some of the targets twice. The gestures where meant to be interpreted by the robot.

The 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.

For every pointing gesture, we calculated the shortest distance between the pointing line and the target position e_d. We also measure the angular deviation e_θ between the actual and the measured pointing direction from the head to the target position.

Pointing Target No.	Distance in m	Height in m	Angle in deg
1	1.41	0	45
2	2	0	90
3	2.24	0	63.43
4	2.24	0	26.57
5	2.24	0.8	26.57
6	2.83	0	45
7	2.83	0.8	45
8	3	0	90
9	3	1.5	90
10	3.04	0	99.46
11	3.04	1.5	80.54
12	3.04	1.5	99.46
13	3.04	2	80.54
14	3.16	0	71.57
15	3.16	2	71.57
16	3.16	2	108.43
17	3.35	0	63.43
18	3.35	1.5	63.43
19	3.35	2	63.43
20	3.61	0	123.69

Table 1: Location of pointing targets: Distance from the test person, height above the floor, and horizontal angle between target and robot, from the test person's point of view.

We split our experiment data into training and test pointing gestures. The training dataset consists of all the 192 gestures of eight subjects. We used the 192 gestures of the remaining eight persons as the test dataset. We train our model on the training dataset and evaluate the accuracy of the learned model on the test dataset. As length-scale parameter and signal standard deviation in GPR we use 1.0.

Fig. 5 shows the distance and angular errors of the different pointing directions and the pointing direction by our

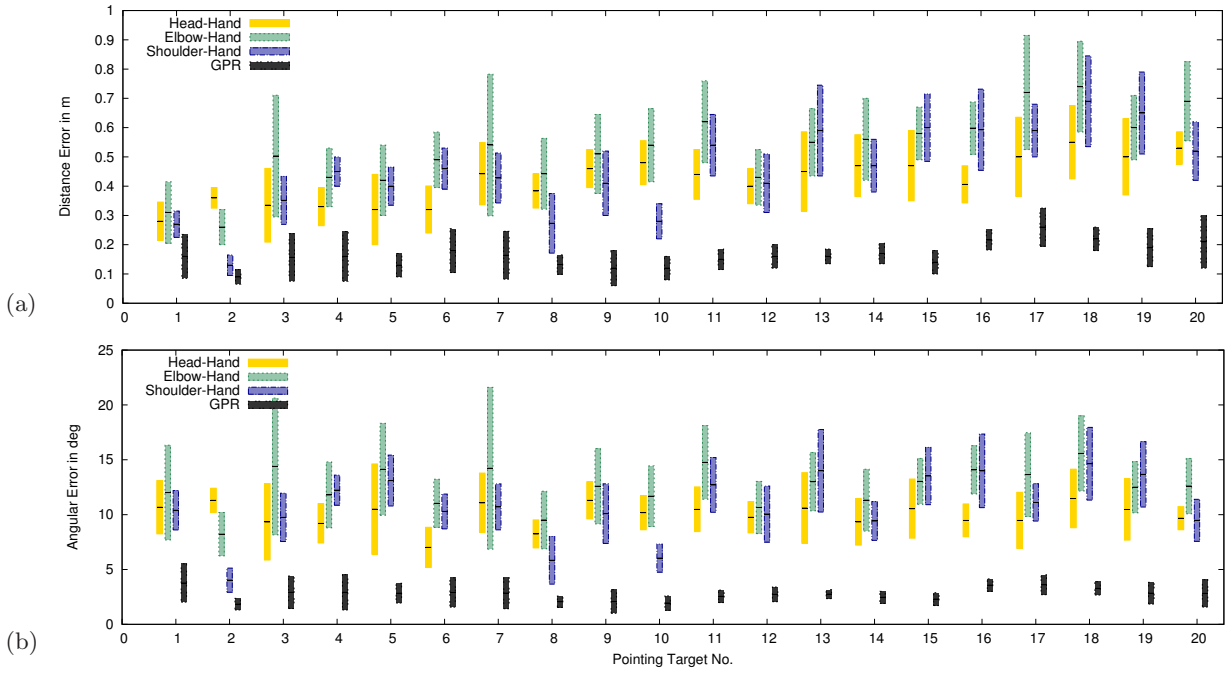

Figure 5: Average distance (a) and angular (b) errors by pointing target for head-hand, elbow-hand, shoulder-hand, and GPR pointing direction. Center and size of the bars indicate mean and standard deviation.

trained model for the pointing gestures in the test data set. The angular errors of the different pointing directions by test person are shown in Fig. 6. The figure includes the head-hand, shoulder-hand and elbow-hand line for the training and test dataset and the pointing direction by our trained model for the test dataset. In addition, angular errors by distance and angle to the target are shown in Fig. 7.

The results show that for many target locations, the head-hand line is a rough but better approximation than the shoulder-hand or elbow-hand line. However, in some situations, the shoulder-hand or elbow-hand line yield lower errors. For targets that are located at an angle of 90° to the right of the persons, the shoulder-hand line yields lower errors in our experiments (cf. Fig. 5).

In all cases, our trained model approximates the pointing direction clearly more accurate than the simple models. Note, that GPR considers the complete arm posture through the low-dimensional gesture representation in the feature space. Our results indicate that simple criteria like the head-hand line are inadequate to estimate the pointing direction accurately, especially in certain ranges of target locations.

One-way analysis of variance (ANOVA) were performed to compare the mean error and standard deviations of the four pointing directions. The ANOVA test showed that there is a significant difference between the mean errors and standard deviations with $F(3, 684) = 108.16, p < 0.01$. Multiple comparisons with the Tukey Honestly Significant Differences (HSD) method showed a significant difference ($p < 0.01$) between the mean errors of the GPR pointing direction and the mean errors of the three other pointing directions.

	e_d in m	σ_d in m	e_θ in deg	σ_θ in deg
Head-Hand	0.39	0.17	9.19	3.94
Elbow-Hand	0.49	0.28	11.88	7.26
Shoulder-Hand	0.44	0.25	10.54	5.56
GPR	0.17	0.12	2.79	1.99

Table 2: Average distance and angular test set error.

The overall average error of all pointing gestures and all test persons is given in Table 2. The table shows that the candidates seem to roughly point in the direction of the head-hand line. Our trained model outperforms the simple estimates clearly.

Compared to the approach by Nickel et al. [15], we achieve a higher accuracy in the pointing target estimation. Nickel et al. find hands through color segmentation and estimate the 3D location of head and hand with a stereo camera system. Their experiment comprises 129 pointing gestures, performed by twelve subjects to eight different pointing targets. The targets have been distributed in a maximum range of 4 m to the camera. They report an average angular error of 25°, using the head-hand line to estimate the pointing direction.

Huber et al. [7] also evaluate the accuracy of their pointing direction estimate. They localize and track body features with a stereo vision system and determine the pointing direction from the shoulder-hand line. They report a mean error of 0.41 m with a standard deviation of 0.17 m. In their experimental setup, two different persons pointed to eight marked positions on the floor. The positions were distributed to the left, to the right, in front, and behind the person with a maximum distance of 3 m in any direction.

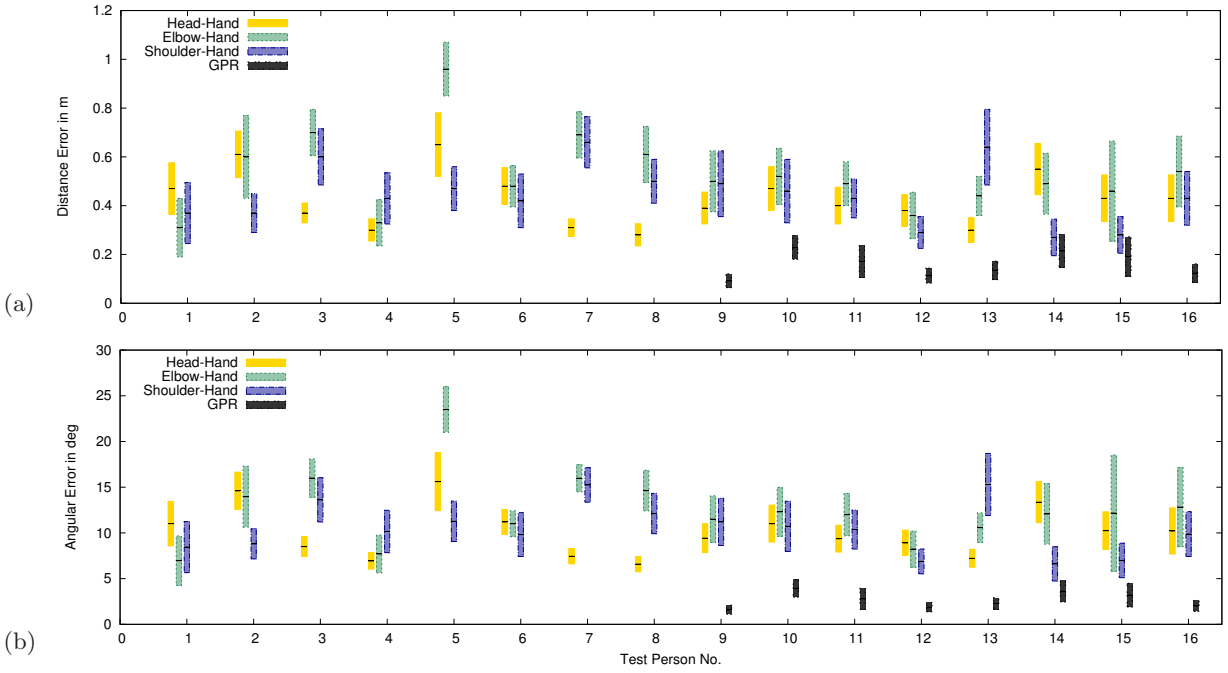

Figure 6: Distance (a) and angular (b) errors by person for the head-hand, elbow-hand, and shoulder-hand direction for test and training datasets. The accuracy of the GPR pointing direction estimate is shown for the test data set (person 9-16). Center and size of the bars indicate mean and standard deviation, respectively.

7. CONCLUSION

In this paper, we proposed an approach to recognize pointing gestures and to estimate the intended pointing direction. Using depth and amplitude images of a Time-of-Flight camera, we localize body parts such as the head, elbow, and hand of a human user. After recognizing pointing gestures, we determine the pointing direction with a model that we learn from human observation. The trained model maps postural features to pointing directions. These features are encoded as angles from head to hand, elbow, and shoulder, relative to the robot's forward direction.

We evaluated the accuracy of the estimated pointing directions in an experiment with sixteen test subjects pointing at twenty different object locations. The results show that our learned model achieves far better accuracy than simple criteria like head-hand, shoulder-hand, or elbow-hand line. Our system also achieves higher accuracy in estimating the pointing direction than reported by Nickel et al. [15] and Huber et al. [7] using a stereo-camera system.

We developed the pointing gesture recognition method for our service robot system. Accuracy is important for this application. The higher the accuracy of the pointing direction estimate, the better can pointing gestures be used to correctly focus the attention of the robot to objects or to disambiguate between objects.

In future work, we will adopt our approach to other parametric gestures, like size-indicating gestures. Our approach is not restricted to amplitude and depth images of a Time-of-Flight camera. The use of color and depth information from an RGB-D sensor and the extraction of more body features could further extend the applicability of our approach.

8. ACKNOWLEDGMENT

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

9. REFERENCES

[1] T. Axenbeck, M. Bennewitz, S. Behnke, and W. Burgard. Recognizing complex, parameterized gestures from monocular image sequences. In *Proceedings of the 8th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2008.

[2] P. Breuer, C. Eckes, and S. Müller. Hand gesture recognition with a novel IR time-of-flight range camera - A pilot study. In *Proceedings of the International Conference on Computer Vision/Computer Graphics Collaboration Techniques (MIRAGE)*, 247–260, 2007.

[3] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[4] F. Faber, M. Bennewitz, C. Eppner, A. Görög, C. Gonsior, D. Joho, M. Schreiber, and S. Behnke. The humanoid museum tour guide Robotinho. In *Proceedings of the 9th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2009.

[5] B. Fransen, V. Morariu, E. Martinson, S. Blisard, M. Marge, S. Thomas, A. Schultz, and D. Perzanowski. Using vision, acoustics, and natural language for disambiguation. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 73-80, 2007.

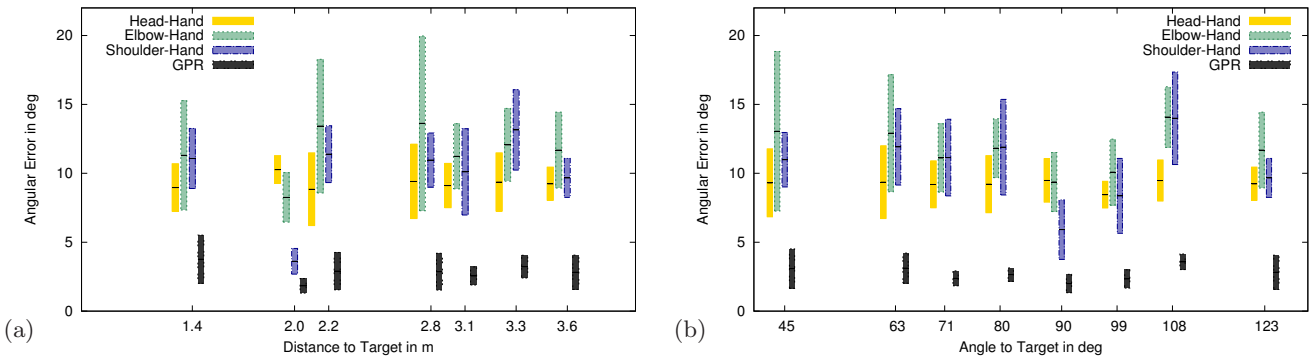

Figure 7: Angular errors for the head-hand, elbow-hand, shoulder-hand, and the GPR pointing direction, by distance to the target (a) and horizontal angle between target and robot, from the test person's point of view (b). Center and size of the bars indicate mean and standard deviation.

[6] M. Haker, M. Böhme, T. Martinetz, and E. Barth. Self-organizing maps for pose estimation with a time-of-flight camera. In *Proceedings of the DAGM Workshop on Dynamic 3D Imaging – Lecture Notes in Computer Science Volume 5742*, 142–153, 2009.

[7] E. Huber and D. Kortenkamp. A behavior-based approach to active stereo vision for mobile robots. *Engineering Applications of Artificial Intelligence*,1998

[8] S. Knoop, S. Vacek, and R. Dillmann. Sensor fusion for 3D human body tracking with an articulated 3D body model. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Orlando, Florida, USA, 2006.

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

[10] M. M. Loper, N. P. Koenig, S. H. Chernova, C. V. Jones, and O. C. Jenkins. Mobile human-robot teaming with environmental tolerance. In *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2009.

[11] 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)*,2007.

[12] 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.

[13] P. McGuire, J. Fritsch, J. J. Steil, F. Röthling, G. A. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter. Multi-modal human-machine communication for instructing robot grasping tasks. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2002.

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

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

[16] J. Stückler and S. Behnke. Integrating indoor mobility, object manipulation, and intuitive interaction for domestic service tasks. In *Proceedings of the 9th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2009.

[17] J. Stückler and S. Behnke. Improving people awareness of service robots by semantic scene knowledge. In *Proceedings of the International RoboCup Symposium*, 2010.

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

[19] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.