Interacting with a Mobile Robot: Evaluating Gestural Object References

Joachim Schmidt¹ and Nils Hofemann¹ and Axel Haasch¹ and Jannik Fritsch¹,² and Gerhard Sagerer¹

¹ Applied Computer Science, Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany
{(jschmidt|nhofeman|ahaasch|sagerer}@techfak.uni-bielefeld.de
² is now with: Honda Research Institute Europe, 63073 Offenbach, Germany
jannik.fritsch@honda-ri.de

Abstract—Creating robots able to interact and cooperate with humans in household environments and everyday life is an emerging topic. Our goal is to facilitate a human-like and intuitive interaction with such robots. Besides verbal interaction, gestures are a fundamental aspect in human-human interaction. One typical usage of interactive gestures is referencing of objects. This paper describes a novel integrated vision system combining different algorithms for pose tracking, gesture detection, and object attention in order to enable a mobile robot to resolve gesture-based object references. Results from the evaluation of the individual algorithms as well as the overall system are presented. A total of 20 minutes of video data collected from four subjects performing almost 500 gestures are evaluated to demonstrate the current performance of the approach as well as the overall success rate of gestural object referencing. This demonstrates that our integrated vision system can serve as the gestural front end that enables an interactive mobile robot to engage in multimodal human-robot interaction.

I. INTRODUCTION

The aim of bringing robots into the household and therefore the need to interact with humans in everyday life is often addressed. But only by adopting to the interaction styles their human partners are used to such robots will be truly personal robots and not technical gadgets. Consequently, research is turning to those perceptual processes that facilitate a human-like and intuitive interaction with robots. Besides verbal interaction, gestures are a fundamental aspect in human-human interaction. One typical usage of interactive gestures that is employed very often when the conversation focuses on the physical environment is referencing of objects. This requires not only the development of individual algorithms for pose tracking, gesture detection, and object attention but also the integration of these different methods in a single vision system. In combination with the analysis of verbal information (i.e., modules for speech understanding and dialog) the visually extracted information allows for situated human-robot interactions involving actions, objects, and references between them.

This paper describes a novel integrated vision system that combines several vision algorithms for enabling a mobile robot to resolve gesture-based object references. The first step is to enable the robot to detect and track the human partner independently of the image background. This is achieved using an articulated 3D body model that is fitted to the image data by applying a particle filtering method. Using a 3D model is more principled than image-space approaches and makes the tracking independent from the viewpoint of the camera. As the next step in the process, the gesture is sought to be identified based on the tracked hand trajectory. A multiple hypothesis approach employing a view-independent gesture representation is used to segment the trajectories and identify the current gesture from a set of learned models. For finally finding the referenced objects, the recognized pointing gestures are used to calculate a region of interest (ROI) representing the estimated object position in the world. Images taken from this ROI could be analyzed with established object detection methods but this is not in the scope of this paper. Rather we will here evaluate the vision front end resulting in the ROI.

Next, we will outline related approaches. Section III illustrates the overall system design and the individual modules integrated. In Section IV we describe the extensive evaluation performed to show the usability and precision of the overall system as well as its parts. The paper ends with a conclusion in Section V.

II. RELATED WORK

Since the early work of Crowley and Coutaz [3], hands and the vision based detection of their motions is known as ways to interact with computer systems. This has developed into complex integrated vision systems, integrating gesture recognition, object detection as well as dialog systems to build up multimodal and interactive systems [1]. Especially the combination of several modalities enables robots to become aware of interactions partners in the real world and to achieve situated and human-like interactions with them. A robot system heading for this goal is presented by Stiefelhagen et al. [16]. In detail they utilize the temporal synchrony of speech and hand trajectories for recognizing pointing actions to objects. However, a specialized depth sensor (stereo or time-of-flight) is needed for tracking the human’s face and hands in 3D, similar to [9] who estimate the direction of pointing using stereo cameras.

But even a monocular camera can be enough for realizing visual recognition and tracking of people and hand gestures, as Germa et al. [5] show with their tour-guide robot. The fusion of multiple visual cues is achieved by a particle
filtering framework, command gestures towards the robot are recognized from hand configuration templates. As Germa et al. neither apply 3D body tracking nor hand trajectory recognition, their scenario does not involve interaction with real-world objects.

In the CoSy project, Kruijff et al. [10] work on aspects of anaphoric and exophoric references to objects. They present a framework to manage the problems of dynamic changes of the environment typical for human-robot interactions. Combining speech and object recognition they achieve learning and detection of objects. However, they do not dwell on gestural input as an additional cue.

Even if each, object recognition, visual attention, language and action processing yield useful results as individual algorithms, Fay et al. [4] prove that integrating sensory data from different modalities is essential for human-like performance in dealing with ambiguities. Brooks and Brazeal [2] present a modular spatial reasoning framework that is able to determine the object referent from deictic reference including pointing gestures from a stereo camera setup and speech.

In our work we aim for an integrated system for resolving object references, where the pointing direction is explicitly recognized based on a tracked 3D body model utilizing a monocular camera and relying on trained gesture models without requiring an explicit triggering of the recognition process. Being part of a bigger framework for a mobile robot, it is convenient for our system to benefit from incorporating verbal information, the robot’s person attention and further modalities. The same framework is also suitable to enable human-oriented interaction with an anthropomorphic robot [15]. However, this paper will focus on the vision aspects only and provide detailed evaluation on the quality of the individual algorithms and the vision system as a whole.

III. SYSTEM SKETCH AND ALGORITHMS

In our study we made use of the mobile robot BIron (Bielefeld Robot Companion [7]). Intended to be a personal robot companion, the robot BIron is a research and evaluation platform used in human-robot studies focused on social interaction. Besides the vision system for resolving object references, the basic functionalities of our robotic interaction system are a model for human awareness, the ability to navigate in unknown environments and a dialog system for naturally spoken language, see also [15].

The vision components presented in this paper are designed to be integrated into a multimodal system that can bear the requirements of human-robot interaction in everyday life.

The contribution of this paper can be seen as part of a bigger learning scenario: Imagine the robot to be placed in an unknown environment, e.g., the first time to see the flat of the human owner. The task here is to let the robot learn new objects by pointing at them, therefore having the human and the robot observing each other and interacting on objects in the real world.

Regarding the technical basis, we use two cameras, one for recognizing the human and his gestures and one active camera for the acquisition of closeup images of objects. For body tracking an Apple iSight (body tracking camera) is used, whose field of view is wide enough to completely observe the upper body of the interaction partner at a distance of 2.0 - 2.5 m. The second camera (object camera) is a Sony Evi pan-tilt camera, utilized to get a closeup of the referenced object. Calibrating both cameras with respect to the robot’s coordinate system allows the object camera to aim at the target position of a pointing gesture seen from the body tracking camera. The software basis for the integrated system presented in this paper is provided by the XML-based communication framework XCF [18] and the image processing framework IceWing [11].

The robot system (cf. right side of Fig. 2) provides information about the environment and the human interaction partner. The person attention processes multimodal cues, including speaker localization as well as face and legs detection for finding and tracking possible interaction partners. The natural spoken dialog system is used for speech input and output; the execution supervisor controls the different modules during interaction. Sensory data and processing results are stored in a scene model and made available for other modules.

Fig. 1. Evaluation scenario: Robot BIron facing a table with objects.

Fig. 2. System sketch for recognizing pointing gestures on a mobile robot.
recognized action, a region of interest (ROI) is determined (Object Attention, Section III-C) to finally enable detecting and/or learning the referenced object.

### A. 3D Body Model Tracking

In the presented scenario we assume the human is standing in front of the robot with the upper body being visible in the gesture camera. For estimating the pose of the human for each frame, we apply the monocular 3D body model tracking algorithm described in [14]. The motions of the person are tracked using the articulated model depicted in Fig. 4 with 14 degrees of freedom. Estimated parameters are the humans position and orientation (each 3 DOF) with respect to the robot and the joint angles of the two arms (3 DOF for each shoulder, 1 DOF for each elbow). Currently, we apply an individual model for each person observed and initialize tracking by hand, assuming that the person attention module provides a coarse positioning.

To match a given pose of the model with the image data, the model is first back-projected into the image, then evaluated for each limb of the model based on multiple color- and intensity-cues. During system design, we deliberately avoided techniques employing background adaptation like silhouette computation, to allow tracking to work in unknown and changing environments or even while the robot or the camera move.

![Fig. 3. Tracking results for subject A while pointing at an object.](image)

Although depth information is not directly observable in a monocular setup, it can be inferred from the best configuration of the articulated model. The pose tracking algorithm thereby provides 3D motion trajectories of the upper body in the robot-centered coordinate system.

### B. Trajectory-based Gesture Recognition

The tracked motions are used as input for the gesture recognition in order to detect meaningful gestures. The use of a 3D model enables the robot to track motions independent from the viewpoint of the camera. Motions are represented in a cylindrical coordinate system with its basis in the human’s shoulder, the features used for the recognition process are the relative radial and the vertical velocities with respect to the torso. In a preprocessing step the sequential data is smoothed by a causal Savitzky-Golay filter [12].

![Fig. 4. Articulated 3D upper body model with 14 DOF.](image)

Due to the constraints of the presented scenario, the trajectory of the right hand is sufficient for recognizing the human’s gestures. Trajectory models are generated from annotated training examples. For the presented approach, we trained five models $m^k$ for the following gestures: pointing at an object (point), retracting the arm after pointing (back), raising the arm (up), waving (wave) and lowering the arm finally (down), referred by $k \in \{p, b, u, w, d\}$.

Briefly, the actual recognition is done by comparing the observed motion $z$ with each of the learned trajectory models $m^k$ seperately, using a time window $[t - I, t]$:

$$d(z, m^k) = \sum_{i=1}^{2} \sum_{j=1}^{t} (z_i(j) - \alpha m_i^k(\phi - \rho_j))^2,$$

with $i$ iterating the dimensions of the feature space.

As a new gesture can start any time, we employ a particle filtering scheme [8] tracking a multitude of hypotheses each with different parameters. Optimized parameters are the type of gesture $k$, the time offset $\phi$ when the gesture started and scaling factors for duration $\rho$ and amplitude $\alpha$ of the motion, respectively. These parameters form the state vector of a particle. Its values define a trajectory hypothesis that is compared with the trajectory actually observed based on the similarity measure $d$. From this comparison the likelihood for each hypothesis can be calculated and used for the propagation step in the particle filter. For each trajectory model $m^k$ the likelihoods of all hypotheses can be added over the complete particle set resulting in the likelihoods for each motion $k \in \{p, b, u, w, d\}$ at any given time (see Fig. 5(a)) and for completing the gesture successfully (see Fig. 5(b)).

The gesture recognition thereby segments the continuous hand trajectory into meaningful gestures. External triggering, e.g., by speech is not needed.

### C. Object Attention

The presented interaction scenario eventually aims at recognizing objects the human points at. The tracked and recognized gestures provide the basis for calculating a region of interest (ROI) to resolve the referenced object. Only gestures that have been recognized as ‘pointing’ are considered for further processing. The position of the human is compared to the robot’s person attention information to ascertain that the tracking results originate from the interaction partner.
IV. EVALUATION

For evaluation the following setup has been used: The human is standing in front of a table with five objects, facing the robot BIRON. The person is asked to show the objects to the robot by pointing at them. The human’s actions are recorded by the body tracking camera while we ensured that the upper body is in its field of view. Likewise, the object positions are well within range of the pan-tilt camera.

As the evaluation was designed to test each system component on its own as well as the system as a whole, ground truth is needed for the different processing steps. All object positions have been measured in the coordinate system of the robot. Likewise the video data has been manually annotated for the evaluation of the body tracking by segmenting and naming every action. The detected pointing gestures are consecutively used to calculate the ROIs and to acquire closeup views of the referenced object for later use and/or training in an object recognition system.

The other half were experienced users. After an initialization phase for the body tracking the person pointed (point) at each object in sequence (crocodile, cup, ball, lemon, bottle), withdrawing their hand (back) after each gesture. The next gestures are raising the hand (up) and waving (wave) into the camera, finally lowering the arm (down). All participants had to perform the same sequence of gestures three times changing the order of targets each time.

We recorded each subject performing the experiments four times, resulting in 16 sequences with a total number of 18572 images, equivalent to more than 20 minutes of video, counting 496 performed gestures in total.

For evaluating the experiments, the 3D body model tracking is manually initialized by determining an initial pose. The resulting hand trajectories of the body tracker are used as input for the gesture recognition. The detected pointing gestures are consecutively used to calculate the ROIs and to acquire closeup images of the objects. In the following we discuss the recognition results of each module in detail.

A. Evaluating Body Tracking

The body tracking algorithm has been evaluated on three of the recorded sequences with 836 images in total. Ground truth has been generated by manually annotating the position of the human’s hands and the head in the images. The measures of the body model and an initial posture are also defined manually for each sequence.

![Fig. 6. Region of interest for a pointing gesture.](image)

To determine the ROI, the pointing direction is assumed as a line extending from the users head over his hand, similar to [16]. The region of interest is represented as a sphere in the world coordinate system, centered at the estimated object position. The distance between the tracked hand position and the ROI center can be scaled comprising verbal information, e.g., describing the size (‘large’, ‘tiny’) of the object. The zoom of the object camera is set accordingly. A fixed distance is used for this evaluation, as speech input has not been considered. The ROI position is then transformed into the coordinate system of the robot’s object camera and an image is captured for analysis, see Fig. 6. The object attention verifies if the verbal and visual information correspond during interaction. Predefined color words, e.g. ‘blue’, are used to apply a simple color based object segmentation using a direct mapping to preassigned RGB-values. This enables the user to resolve ambiguities from pointing, e.g., for two nearby objects by naming their color. Thus a coupling of verbal and gestural modalities is supported by the object attention. The color segmentation can be used to acquire a closeup view of the referenced object for later use and/or training in an object recognition system.

![Fig. 7. Tracking error of subject B for two configurations. Mean errors µ as in Tab. I. Pointing at object (5) around frame 270 is difficult to track.](image)

We configured the system in two ways: 1) Using 1500 particles and 6 meanshift iterations takes approx. 5 sec/image for computation. This configuration is used for carrying out the overall system evaluation. 2) Using 500 particles and three meanshift iterations results in a lower accuracy but achieves faster computation at approx. 1-2 Hz, shown for comparison only. Tab. I shows that we still need a relatively high number of particles to ensure accurate tracking over a longer image sequence. Gestures that effect in self-occlusion...
are more likely to produce noisy results (cf. Fig. 7 around frame 270), as the observation function for occluded limbs, e.g., the torso provides imprecise or even wrong results. Fusing multiple cues for each limb and considering the body model as a whole helps to keep up tracking even if detectors for individual body parts were mislead. Both evaluated configurations are able to recover from tracking errors, but only 1) permanently provides reliable trajectories.

B. Evaluating Gesture Recognition

The gesture recognition is evaluated using the trajectories provided by the body tracking component. The models of the gestures are trained separately for each person using the tracked and annotated motions from three of the four runs. The remaining sequence is used for evaluation. This is resulting in sets of only five to 20 training samples for each gesture model. For the intended application, fast adaption to new users is an important feature but obviously prevents extensive collection of training data. The presented algorithm shows robust performance even if only a low number of training examples are provided. Note, that systematic errors of the body tracking are co-learned in the training process, thus enabling robust gesture detection even for coarse tracking results.

Table II is illustrating the recognition results and error types for the five different motions. In each column the number of annotated and correctly detected gestures is noted followed by the error types of double detections, deletions, false insertions and substitutions. As well the resulting error and recognition rates are given. The last column is summarizing the results. Subject IV executed the wave motion simply by moving the hand instead of the forearm which is not recognizable for the body tracking. The marked column (wave*) therefore displays the results neglecting subject IV for a better comparability.

![Fig. 8. ROI center positions for recognized pointing gestures and ground truth (bold markers) for each object. Objects (1-5) from left to right.](image)

The gesture recognition shows a high recognition rate of 94.6% and an acceptable low error rate of 17.3%. This rate is calculated by adding up all errors (deletions, insertions and substitutions) for the respective gesture. The false positive recognitions (insertions) make up the major part in this error (11.9%; e=59). For a robot system, this could lead to false assumptions about the gestures performed by the user. Compared to our previous work [6], appropriate depth information about the current body pose can significantly improve the robustness of action recognition. Probabilistic fusion of data from gesture and speech recognition, however, could help to validate assumptions, as speech and gestures naturally co-occur in everyday communication [17].

C. Evaluating Object Attention

Both, body tracking trajectories and recognized gestures are essential for determining the region of interest. For evaluating the object attention module only the 167 correctly recognized pointing gestures are used. Five of those gestures have been neglected by the object attention (cf. Table II), because their calculated ROI was outside the interaction space observable by the object camera. Furthermore, no verbal information is used, so that the distance between the ROI and the hand is not altered, as the color based selection of objects and the segmentation of the ROI image is not in the scope of this paper.

The final object position error is calculated as the euclidian distance in [m] in the world coordinate system between the measured object position and the ROI center. Table IV-B presents the RMSE error $\mu$ and the variance $\sigma$ for the ROIs separately for each subject. The objects (1-5) are ordered in the same manner as in Figures 8 and 3.

D. System Performance

The ROI errors are for most objects not bigger than 25cm. The error in the ROI positions is mainly due to the coarse gesture information from the preceding modules. Also, estimating the pointing direction as a line from the head through the hand does not always suit well. For objects (4) and (5) to the left of the person (at the right in the image), the ROI is estimated to be too close to the human (cf. Fig. 8) for two reasons: 1) The body tracking cannot handle the self occlusion that good and produces more noisy results (cf. Fig. 7), 2) for the described posture it is uncommon to have the hand in the line of sight towards the object, the gesture is more like ‘shooting from the hip’. This calls for learning task-specific dependencies. The error could be overcome by altering the ROI offset for this specific gesture, as the recognized ROI positions are still compact (low $\sigma$ in Table IV-B) but their mean position is slightly offset (cf. Fig. 8).

Objects (1-3) show that better results are achievable. Comparing the different subjects, the errors vary only slightly
although the gestures were performed very differently across subjects. Individual gesture models therefore seem to be a good way to deal with such variations.

V. CONCLUSION AND OUTLOOK

In this paper we presented a short overview and an extensive offline evaluation of the vision system performing the processing of visual input on our mobile robot BIRON. Employing only the visual input channel, the quality of understanding object references is already quite good. The exactness could be further increased if verbal descriptions for size and color were available and information from the person attention (e.g., position of the gesturer, acoustic input present) were utilized. Integrating context knowledge about the environment and the current task could also help to reduce the search space for the body tracking as well as the gesture recognition. The 3D body tracking is the only part of our vision system that does not yet provide a robust tracking when configured to work in real-time. Calculation speed could be improved by restricting the search space through the learned motion models or by refining the observation model to provide accurate results for even less cues. Automatic initialization and error recovery of the tracking algorithm [13] will help to make the system more usable for human-robot interaction. Ultimately, the presented system will be the basis for a sophisticated human-robot interaction system. As the body tracking operates background-independent and a single camera would be sufficient, the system is well suited for application on a humanoid robot, contributing to more human-like interaction with such a robot.

REFERENCES


<table>
<thead>
<tr>
<th>Object</th>
<th>subj A</th>
<th>subj B</th>
<th>subj C</th>
<th>subj D</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># g</td>
<td>µ</td>
<td>σ</td>
<td># g</td>
<td>µ</td>
</tr>
<tr>
<td>(1) green crocodile</td>
<td>9</td>
<td>0.22</td>
<td>0.04</td>
<td>7</td>
<td>0.10</td>
</tr>
<tr>
<td>(2) blue cup</td>
<td>8</td>
<td>0.28</td>
<td>0.02</td>
<td>10</td>
<td>0.21</td>
</tr>
<tr>
<td>(3) pink ball</td>
<td>8</td>
<td>0.29</td>
<td>0.06</td>
<td>7</td>
<td>0.14</td>
</tr>
<tr>
<td>(4) yellow lemon</td>
<td>7</td>
<td>0.38</td>
<td>0.05</td>
<td>7</td>
<td>0.25</td>
</tr>
<tr>
<td>(5) red bottle</td>
<td>8</td>
<td>0.58</td>
<td>0.04</td>
<td>10</td>
<td>0.47</td>
</tr>
<tr>
<td>all</td>
<td>40</td>
<td>0.35</td>
<td>0.14</td>
<td>41</td>
<td>0.25</td>
</tr>
</tbody>
</table>

TABLE III

Position error for the calculated region of interest (ROI) for all objects (1-5) and subjects (A-D). # g: number of recognized gestures, µ: mean error, σ: standard deviation, both in [mm].