Spatial References with Gaze and Pointing in Shared Space of Humans and Robots

Patrick Renner[1], Thies Pfeiffer[2], and Ipke Wachsmuth[1]

[1] Artificial Intelligence Group
[2] Cognitive Interaction Technology Center of Excellence
Bielefeld University
Universitätsstr. 25, 33615 Bielefeld, Germany

Abstract. For solving tasks cooperatively in close interaction with humans, robots need to have timely updated spatial representations. However, perceptual information about the current position of interaction partners is often late. If robots could anticipate the targets of upcoming manual actions, such as pointing gestures, they would have more time to physically react to human movements and could consider prospective space allocations in their planning.
Many findings support a close eye-hand coordination in humans which could be used to predict gestures by observing eye gaze. However, effects vary strongly with the context of the interaction. We collect evidence of eye-hand coordination in a natural route planning scenario in which two agents interact over a map on a table. In particular, we are interested if fixations can predict pointing targets and how target distances affect the interlocutor's pointing behavior. We present an automatic method combining marker tracking and 3D modeling that provides eye and gesture measurements in real-time.

Keywords: shared space, human-human experiment, gaze tracking, gesture prediction, automatic interaction analysis

1 Motivation and Overview

The way we interact with robots changes more and more from simple task instructions to cooperative settings with a close interaction between humans and robots in a shared space. If the peripersonal spaces of the interaction partners overlap, they form an interaction space [19] in which actions need to be well coordinated to avoid harm and to ensure a successful and swift task completion. This raises completely new requirements regarding a robot's skills in interacting, especially considering the timing of actions and space allocations.

For interaction in shared space, a robot needs to be aware of its immediate surroundings. A dynamic representation of the robot's peripersonal space can be used to detect human arm movements (e.g. [12]) and thus prevent collisions by stopping the robot's movement, e.g. when both human and robot want to point to a certain target.

There are several levels on which interaction with a robot could be improved if the robot were able to follow human eye gaze and to predict upcoming human gestures: First, current action executions could be slowed down or halted if the robot notices that the human's intended movements would conflict with its current target. Second, in a more proactive manner, the robot could turn its head towards the predicted target, which would serve several functions: It would communicate the robot's interpretation of the human's movements to the human interlocutor and by this means facilitate communication robustness and increase the confidence of the human in the grounding of the current target [5]. In addition, sensors attached to the robot's head could be timely oriented towards the expected target to collect optimal data on the human gesture and its interaction with the target. Third, anticipated gesture trajectories could be considered during action planning to avoid potentially occupied areas.

So long, the focus has been on the interpretation and anticipation of human actions by the robot. For a successful interaction, the robot should also be enabled to provide signals that can in turn be used by the human interlocutor to make similar anticipations. Staudte and Crocker [28] showed that a human-like gaze behavior of a robot can positively influence human comprehension of robot speech and thus improve the interaction.

Hence, a better understanding of human skills for anticipating movements could help robots to increase robustness and smoothness of shared-space interactions. As a starting point for this idea, we investigated the coordination of gaze and gesture (see e.g. Fig. 2) in a human-human study using a complex task. In addition, we assessed the coordination of pointing gestures and upper body movements. For example, to reach far targets, humans will need to lean forward. A prediction model considering both gaze and gestures could enable the robot to have a detailed concept of a starting human pointing movement.

In the study reported in this paper, gaze and pointing directions as well as the head positions of the participants were recorded in a route planning scenario (motivated by former work of Holthaus et al. [11], see Fig. 1). Analyzing mobile eye tracking data usually requires manual annotation which would render a human-human study tedious if not unfeasible due to the needed effort in time. In addition to that, manual annotations are less precise and thus do not provide sufficient data for model generation. Therefore, an automatic method was developed combining fiducial marker tracking and 3D-modeling of stimuli in virtual reality as proxies for intersection testing between the calculated line of sight and the real objects (see also [23]). The method merely relies on the scene camera video of the mobile eye tracking device for mapping eye gaze on targets in the 3D environment. For the study, this set-up was extended with an external tracking system for recording pointing gestures. In the future target scenario of human-robot interaction, tracking of the hands could be done with the tracking sensors mounted on the robot.

The remainder of this paper is organized as follows: After discussing related work in section 2, an experiment on spatial references with gaze and pointing in a route planning scenario is reported in section 3. A novel method for auto-

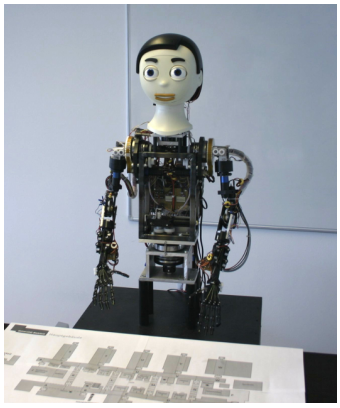

Fig. 1. The route planning task of the present study is motivated by a receptionist scenario [11, p. 4].

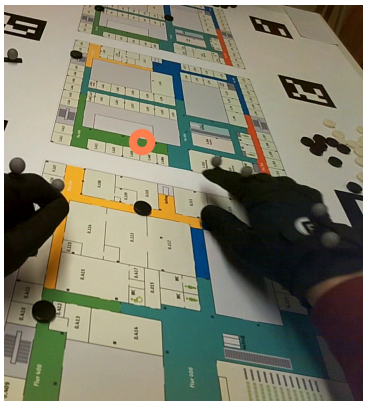

Fig. 2. Example for a fixation (highlighted by the ring) anticipating the pointing target.

matic analysis of the acquired data is proposed in section 4. The results of the experiment are presented in section 5 and discussed in section 6.

2 Related work

The prediction strategies investigated in this paper focus on non-verbal behavior in shared space. We will thus first discuss representations of peripersonal and shared space before we attend to different strategies making use of such representations that have already been implemented on robots. We will finally present findings on human-human interactions regarding the coordination of gaze and pointing gestures depicting insights which motivate our approach to predicting pointing targets in particular and target areas of movements in general.

2.1 Human representation of space

If we want to reach an object, we instinctively know if we are able to do so without moving our torso and we know how far to reach. There are different explanations for these skills. Rizzolatti et al. [27] and follow-up research on humans [25] suggest that the space immediately around oneself has an own, specific neural representation. This space is called peripersonal space. Objects in peripersonal space can be reached without moving the torso. The neural representation of peripersonal space integrates visual, tactile and even auditory signals [10][7]. It allows for constant monitoring of the position of objects in that space relative to the body. Clinical studies by Làdavas [15] show that peripersonal space can adjust with the position of the body parts. It is even possible to enlarge it by grasping objects and using them as tools [3].

For the conceptualization of shared space, Kendon [14] proposes an activity space for each partner, similar to the peripersonal space. The overlap of the partners' activity spaces then forms a common space for interacting: the O-space. Nguyen and Wachsmuth [20] combined Kendon's O-space with the peripersonal space, defining interaction space as the overlap of two peripersonal spaces. Moreover, they propose a process of spatial perspective taking for estimating the extent of the partner's peripersonal space in a virtual human.

Humans are well capable of estimating if targets are in reach, by taking into account not only the plain distance to a target, but also the surface layout on which it is located [6]. Hadjidimikratis et al. [9] found evidence for a neural process to evaluate the 3D distance between the eyes and objects: According to their findings, a neural representation of the peripersonal space related to the eye position is used to compute if objects are reachable.

Mark et al. [16] found that people do not try to reach an object without bending the torso until absolutely necessary (the absolute critical boundary), but instead have a preferred critical boundary for starting to lean forward. The preferred critical boundary is thus the point from which reaching is more comfortable when supported by bending the torso. It appears to occur from 85% of an absolute critical boundary. Leaning-forward is also an often used strategy when pointing to distant objects to increase pointing accuracy [22]. Based on these findings, Nguyen and Wachsmuth [20] introduce the lean-forward space to model the gradual transition between peripersonal and extrapersonal space.

Some of these principles have already been implemented to improve robots' capability to flawlessly interact with humans in shared space, as we discuss in the following section.

2.2 Human-robot interaction in shared space

Humanoid robots can be assumed to be anthropomorphized by humans. Thus, it is meaningful for a robot to understand and make use of human behaviors and expectations. Hüttenrauch et al. [13] compared distances participants maintained from a robot in accordance with Hall's theory of proxemics and Kendon's F-formations [14]. In a Wizard-of-Oz study where participants had to show around the robot in a home-like environment, they found that the personal distance is predominant and the vis-a-vis formation is preferred in most interactions. Mumm and Mutlu [17] observed humans' proxemic behavior when interacting with a robot. They manipulated likeability and gaze behavior of the robot. In the dislikeable condition, participants increased their distance to the robot when the robot showed increased eye contact. This suggests a coupling between those different factors for proxemic behavior. Spatial prompting is another proxemic behavior which robots may use in the gaps between two consecutive interactions [8]: By giving subtle cues, a robot can positively influence the spatial positioning of the user.

Concerning interaction in shared space, Antonelli et al. [2] propose an implicit representation of peripersonal space by experiments of gazing and reaching. This way, the robot learns a visuomotor awareness of its surrounding without making

the representation explicit. An explicit approach was developed for the virtual agent Max [19]: After learning the body structure utilizing virtual touch and proprioception sensors, the dimensions of Max's peripersonal space are calculated. It is divided into a touch space, a lean-forward space and a visual attention space. The interaction space is established by spatial perspective taking, i.e. projecting the agent's own body structure onto the partner. Using the findings of an experiment conducting humans' expectations of grasping decisions of the iCub robot, Holthaus et al. [12] proposed the *active peripersonal space*, a spatial model covering handedness, distance-awareness, awareness of the interlocutor's actions and moreover accounting for attention and occupancy. The active peripersonal space is represented by a body-centered spherical coordinate system. The model allows for monitoring the overlap with the partner's peripersonal space. Thus, an interaction space can be formed.

If such a model could be extended to include predictions about human actions, the robot could include these in its own planning and thus execute actions in a foresighted way. Therefore, the use of gaze and pointing in human interaction has to be taken into consideration.

2.3 Gaze and pointing in human-human interaction

The eyes are our fastest moving body parts. Due to their dual use as sensor and communication device, they often rest on objects we are planning to use or to refer to. When conducting manual interactions, we need to have a spatial representation of the target and thus gaze is quite naturally linked to hand movements, such as pointing gestures. Once a spatial representation has been built, gaze might no longer be needed to control movements towards a target, but it is still relevant for fine controlled end positioning [1]. In our own work, we found that pointing directions can be determined most precisely when considering the dominant eye aiming over the pointing finger tip [21].

There is evidence for a temporal coordination between gaze and gestures. For example, Prablanc et al. [24] found that hand movements are initiated about 100 ms after the first saccade to the target in a pointing task. However, as for example Abrams et al. [1] argue, such findings are likely to depend on the task and the way stimuli are presented. Thus findings in laboratory conditions might not scale to situations found in natural interaction.

Neggers and Bekkering [18] report a close coupling of gaze and gesture movements. In their study, participants had to point at a first target and then as soon as possible look at a second one, which was presented a bit later. Their participants were not able to fixate the second target while still reaching towards the first and they explain that by a neural inhibition of saccades during manual pointing.

Several contextual factors may influence the interaction between gaze and gesture. Biguer and Prablanc [4] investigated latency shifts depending on target distances: With increasing distance, the first saccade to the target was produced significantly later. However, the corresponding head movement started earlier. Interestingly, the timing of arm movements was not affected by target distance.

3 An experiment on the interaction of gaze and pointing

Following our goal of improving the precision of the spatial representations of current and future actions, we conducted an experiment on eye-hand coordination. We aim at improving the prediction of pointing targets, and thereby the prediction of areas which might be entered by a pointing hand in the near future. The selection of pointing gestures is due to our scenario, but similar effects are expected and have been shown, e.g., for grasping movements. Our main hypothesis and follow-up questions are as follows:

1. Given the onset of a hand movement, the target of a pointing gesture can be predicted by looking at the location of preceding fixations.
2. How accurate and how precise can the target area be predicted?
3. How large is the advantage in timing that can be gained?
4. What influence does the distance of the target have on the performance of the pointing gesture?

As argued above, we use a relatively natural interaction scenario of two interacting, non-confederate participants, instead of a rigid experimental design with a high level of control. In our route planning scenario, the two interlocutors are sitting at a table facing each other. Placed between them is a map on which they perform several joint planning tasks. To record eye movements, we equipped one participant with a mobile eye tracker. This participant is in the following called P1, the interlocutor without the eye tracking system is called P2.

To elicit spatial references to their own peripersonal space as well as to the space shared by both participants, we use three different floor plans (Fig. 3): ground floor, first and second floor. They are placed between the participants so that there is one within each peripersonal space. The middle floor plan is placed in the interaction space. The scenario is designed to yield a lively interaction facilitating frequent pointing gestures to enable joint planning of routes. It has been verified in a small pilot study.

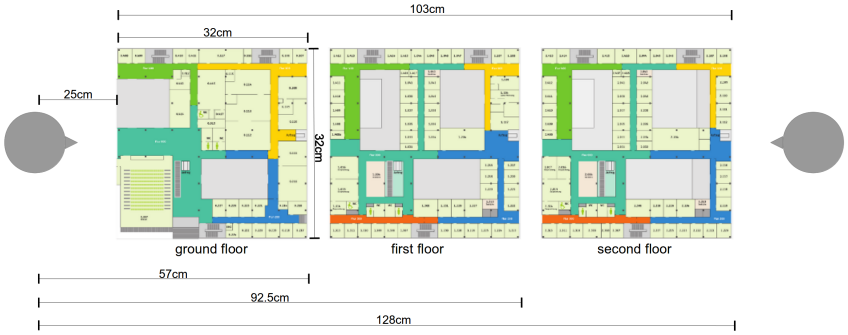

Fig. 3. The figure shows the arrangement of the two participants (left and right, facing each other, sitting) and the three floor plans of the target building on the table between them. The legend shows the relative distances of the different maps to the participants.

3.1 Setup

The three floor plans are printed on a DIN A0 format poster (84.1x118.9 cm), each plan with a size of 32x32 cm. The gap between two plans is 3.5 cm. The poster lies on a table of 130 cm length. Fig. 3 shows the distances to the ends of each floor plan, assuming participants are sitting about 25 cm away from the closest floor plan, in a comfortable position at the smaller sides of the table. By our design, the closest floor plan and the beginning of the middle plan are located in the peripersonal space of the interaction partners. The end of the middle floor plan and the beginning of the far plan can only be reached by more or less articulately leaning forward, depending on the height of the participant. However, the end of the far floor plan cannot be reached while sitting by participants of normal height. The shared space hence comprises the middle plan and the adjacent beginnings of the other two plans.

3.2 Task

In general, the participants' task is to plan routes from a point A to a point B distributed over the three floor plans. Instructions for individual routes are printed on cards with miniature floor plans the participants draw for each stage. Fig. 4 shows the four important steps of the first task type: (a) The task begins with P1 drawing a card where starting point and target room are marked.

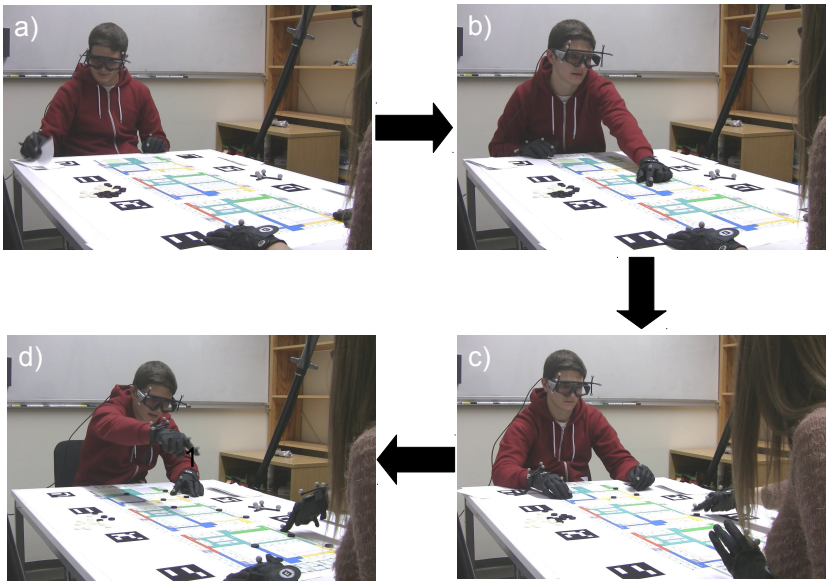

Fig. 4. The main steps of the first task type: (a) Drawing a subtask card, (b) explaining the route, (c) placing blockages and (d) jointly planning the remaining route.

Fig. 5. SMI Eye Tracking Glasses, a binocular mobile eye tracker.

Fig. 6. Gloves with tracking markers attached.

(b) P1 then demonstrates these to the interlocutor and describes the fastest route. (c) The interlocutor P2 then draws a card with blockages and indicates them on the floor plans using gaming tokens. (d) Finally, P1 and P2 jointly plan the remaining route.

This basic task is followed by a second type of task in which complexity is increased: Participants have to plan the fastest route to three rooms at a time. In addition to that, the number of blockages is also increased. In total, each pair of participants had to perform eight repetitions of the first type of task and two repetitions of the second type. The roles of P1 and P2 in the tasks were switched every second repetition.

3.3 Recorded data

During the experiments we collected multimodal data: Two video cameras observed the participants during their interactions. As already explained above, Participant P1 was equipped with mobile eye tracking glasses (SMI Eye Tracking Glasses Version 1.0, 30 Hz gaze data resolution, Fig. 5) to record binocular eye movements and the field of view of P1 using the HD scene camera. This particular eye tracker has parallax compensation and is thus accurate at all relevant distances.

For a precise tracking of the index finger positions of both hands of both participants, an optical tracking system (Advanced Realtime Tracking TrackPack 2) was used. Based on experiences from previous studies we attached the tracking markers for the index fingers to soft golfing gloves which do not compromise the hands' freedom of action (Fig. 6).

4 Method of Analysis

For analysing the temporal sequences of P1's visual attention on the objects of interest while at the same time providing as much freedom to move as possible, mobile eye tracking was used instead of remote eye tracking systems. This choice,

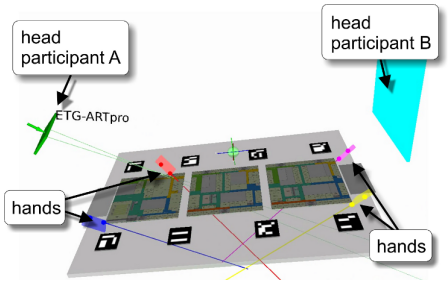

Fig. 7. The 3D representation of the scenario including three floor plans, the fiducial markers, participants' hands with highlighted pointing direction as well as the positions of the interlocutors' faces.

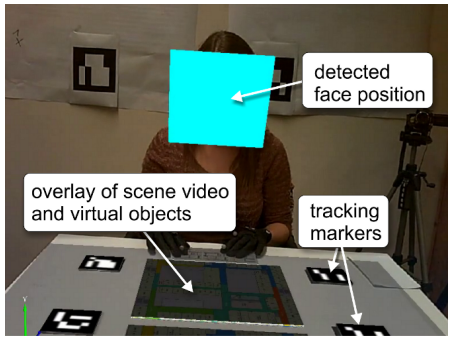

Fig. 8. Overlay of the virtual representation over the scene camera image of the eye tracking device. The position of the participant's face is calculated by a face detection algorithm and automatically masked.

however, normally comes along with time-consuming manual annotation of gaze videos: In the present study, the duration of a recorded sessions ranged within 20-30 min, which would have to be annotated frame-by-frame.

For a different previous study, we designed the EyeSee3D approach [23]. It allows us to automatically assign fixations to stimuli based solely on the scene camera video and the 3D gaze vector provided by the eye tracker. This is done by tracking simple fiducial markers in the video. If the scene camera is calibrated, i.e. its intrinsic parameters are determined, marker positions found in the 2D images can be correctly transformed to 3D with respect to position and orientation. In other words: We can calculate the pose of the scene camera, and thus the head of participant P1, with respect to the stimuli. By re-modeling the complete scenario using virtual reality technology and representing the 3D gaze vector as a 3D ray in space, fixations can be automatically assigned to the stimuli by geometric intersection testing. The 3D model can be inspected from all perspectives (Fig. 7), or it can be aligned to the scene camera video (Fig. 8). This way, the experimenter can use EyeSee3D to monitor the recording process and validate data quality during runtime.

For the present study, we had to extend the EyeSee3D approach in several ways: In addition to using gaze directions, the external tracking system was integrated to accurately detect pointing gestures. The link between marker tracking and external tracking is established by placing a tracking target with known position and orientation relative to the fiducial markers (see Fig. 9(a), next to the middle floor plan). This way, both inputs can be fused in a multimodal 3D representation of the interaction scenario.

The floor plans contain a high number of rooms. Instead of modeling each of these separately (which normally would be necessary as each room is a stimulus) as 3D objects, a technique from web design was made use of: The plans were

represented as webpages with HTML image maps mapped on the surface of the 3D table model. Each annotated area of the image map represents a room or floor and can be tagged with a specific text that is output when the user fixates the area on the floor plan. This approach reduces modeling effort and increases system performance.

We are also interested in fixations on the interlocutor's face, but in this free kind of interaction the partners show large head movements. A static representation is thus not reliable. To overcome this issue, we adapt dynamic position changes by detecting faces in the scene camera images of the mobile eye tracker using the Viola/Jones algorithm [30]. To complete the 3D model, the face position is approximated in 3D as well, which can be seen in Fig. 7 and 8.

5 Results

The study was conducted with pairs of 18 (9 male, 9 female) participants of an age between 18 and 27 (23.5 years on average; standard deviation 2.9 years). From these, seven male and two female participants took the role of P1 by wearing the eye tracker. Which participant of a pair got to wear the device was chosen with regard to visual disorders: The utilized eye tracker is difficult to calibrate with participants wearing glasses, thus usually participants without glasses were preferred for wearing the eye tracker. The arm length (measured from the head to the fingertip) of the participants P1 ranged between approximately 65 cm to 70 cm. In total, 338.8 minutes (or 5.65 hours) of interaction were recorded during the experiments.

After the first five pairs of participants, the setup was mirror-inverted to prevent possible influences of the different floor plans on the results. So, the ground floor was located in front of P1 for the pairs one to five, and in front of P2 for pairs six to nine. For abbreviation, the first pairs will be called NT, the pairs with mirrored setup MT.

In the course of the complex tasks, participants had to visit all three floor plans. Fig. 9 shows a typical course of interaction of task 2: Both participants point to locations near themselves and to the middle floor plan. Communicative fixations – fixations longer than 500 ms [29] – occur in areas near the pointing targets. In total, 642 pointing gestures were conducted by P1 and 681 by P2. In the NT, more gestures were performed by P1 which changed in the MT. Thus there is a slight shift of focus to the ground floor. For pointing, the average coordinates (in meters) over all targets on the table were $(0.002, -0.007)$, close to the center of the plan, which testifies that our targets were well distributed over the area. As Fig. 9 shows, from the view of a participant, the x-axis describes left and right, the y-axis the distance to the middle of the table (coordinates $(0, 0)$). For placing blockage tokens the average coordinates were $(0.007, -0.009)$. Thus, regarding both, the targets of pointing and of placing tokens, interaction on the floor plans was symmetrical between interlocutors.

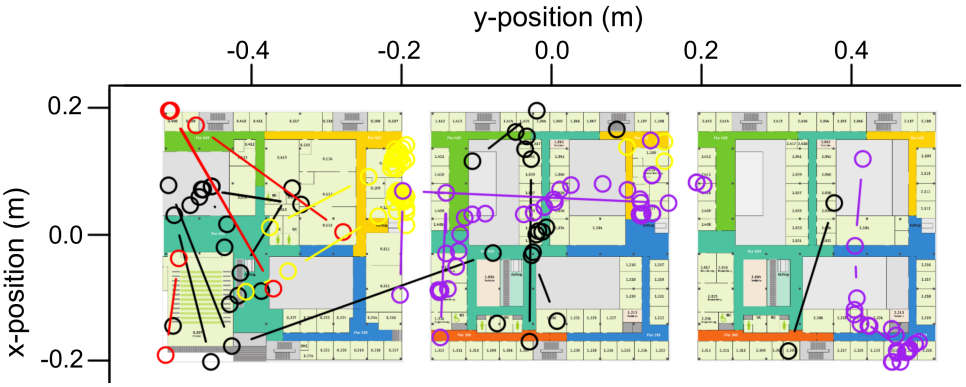

Fig. 9. A typical interaction in a route planning task: Communicative fixations are inscribed in black, pointing directions of P1 in red and yellow, those of P2 in blue and purple. The lines represent events that occurred after one another.

5.1 Gaze-pointing coordination

We first measured when and where fixations were made before the apex of a pointing gesture (stroke). In particular we measured whether these fixations targeted the area around the pointing target. For describing the area we used a circle with a radius of 10 cm. For each participant P1, Fig. 10 shows the percentage of pointing gestures in which at least one fixation hit the target area starting from 10 seconds before the apex onwards in bins of 1 second. As expected, fixations preceding one's own pointing gestures were likely to occur in a small area around the pointing target. For all except two participants we observed an increased percentage of fixations on the target from 2.5 s before the apex of the gesture on; one second before the apex, the percentage of fixations on the target was already higher than 50%. A similar trend can be seen for fixations of P1 on pointing targets of P2 (Fig. 11). In both cases two pairs of participants did not adhere to this scheme. As performing a pointing gesture took on average 511 ms (sd: 225 ms) from onset to stroke, gaze could be a hint for the target as soon as the hand starts moving. When averaging over the fixations during the last 200 ms before the gesture onset, at the time of the onset, the target area could be predicted in 47.7% of all pointing gestures performed; 51.1% when taking into consideration the 200 ms around (i.e. from 100 ms before to 100 ms after) the onset. Extending the target radius to 20 cm around the pointing target, 74.2% of pointing targets could be predicted about 500 ms in advance of the apex of the gesture.

Also the probability of fixating the target of a pointing gesture of the interlocutor raised when closer to the onset of the gesture. However, it is not sure whether participants actually predicted pointing targets of their interaction partner, or if this finding is due to context from e.g. the spoken conversation.

Fig. 12 illustrates the observed fixations on the pointing target of participants P1 before the stroke of the gesture. While fixations were widely spread around the target 2000 ms before the apex, 500 ms before the apex the target area was most frequently fixated.

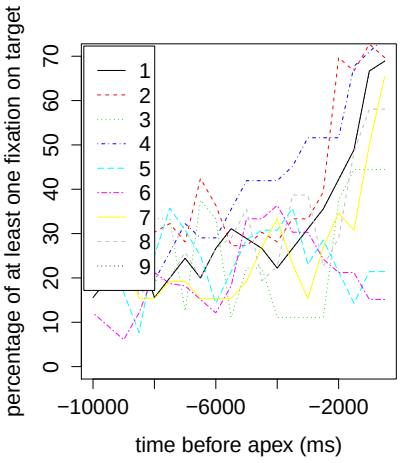

Fig. 10. The probability of fixations hitting the target area of one's pointing gesture increases the closer to the gesture's apex.

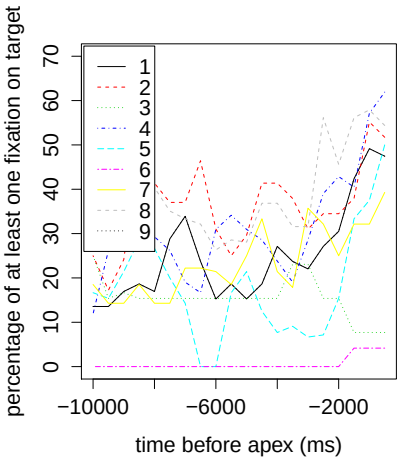

Fig. 11. This also holds when observing the pointing gesture of an interlocutor but not as strongly.

5.2 Characteristics of pointing gestures

By considering the head positions of P1 (the 3D head position of P2 was not acquired) during the stroke of the pointing gestures, body movements could be measured as well. In particular, leaning forward was investigated. Fig. 13 shows the y-position (i.e. the position in direction of the interlocutor) of the fingertip during the apex of a pointing gesture in relation to the head position in y-direction. The red curve is the approximation of the points by locally weighted polynomial regression. When the fingertip in a pointing gesture was not moved further than to the middle of the table, head movements were barely observed. Beyond that distance, leaning forward was performed for pointing. This might be due the distance of 65 cm from the edge to the middle of the table, which could mark the critical boundary from where leaning forward is preferred according to Mark et al. [16]. A trend to leaning forward was already observable from the beginning of the middle floor plan. The same trend, but less distinct, was found in the correlation of head position and the pointing targets (Fig. 14). Fig. 15 shows that, depending on how far the pointing target was away, there was a trend of not directly touching the target but rather pointing from a higher position in distance.

Summing up, most pointing gestures were performed close to the participant pointing. Thus, participants preferred pointing in their own peripersonal space, but the overall distribution of pointing targets of both interaction partners was similar over all floor plans.

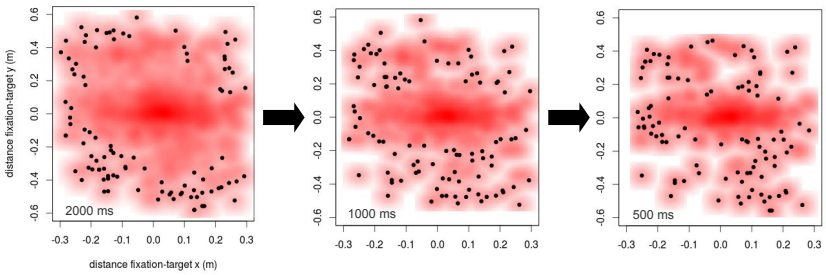

Fig. 12. Focussing the pointing target: From left to right, the distribution of all fixations (relative to the pointing target) up to the pointing gesture apex are shown for 2000 ms, 1000 ms and 500 ms (each for an interval of the next 1000 ms), averaged over all participants.

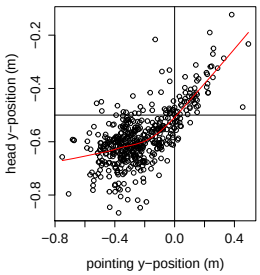

Fig. 13. Correlation between the apex position of the fingertip and the head position in pointing gestures.

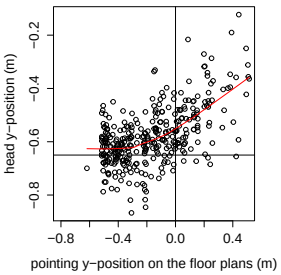

Fig. 14. Correlation between the target position of pointing gestures and the head position.

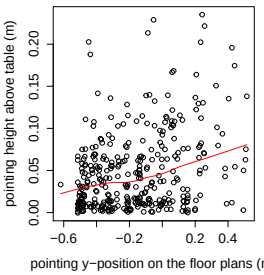

Fig. 15. The height of pointing gestures dependent on the target on the floor plans.

Measuring the target position during the apex of pointing gestures, 66-73% of all gestures targeted locations in the closer half of the table, 44%-52% to the closest plan and only 10-15% to the most distant floor plan. Fig. 16 shows the estimated kernel densities for the fingertip positions at the apex of pointing gestures and the locations pointed at. The gaps between the floor plans are reflected in the graphs. There is a continuous decrease of gestures to targets beyond the middle of the table. The percentages for the position of the finger tip at the apex of pointing gestures were as well similar in all trials and conditions. Here, the NT and MT are combined. In 83% (82.5% for P1 and 82.6% for P2) of all pointing gestures, the hand remained in the own half of the table, i.e. in peripersonal space. In 64.2% (P1) and 59.2% it was within the extent of the nearest floor plan, in 96.4% (P1) and 93.7%, it was located not farther than the end of the middle floor plan.

Thus, a distribution can be observed where pointing movements were performed: A high percentage of pointing targets were located in near space, thus the first floor plan. From the middle of the table, when leaning forward had to be used, there was a continuous decrease of conducted pointing gestures.

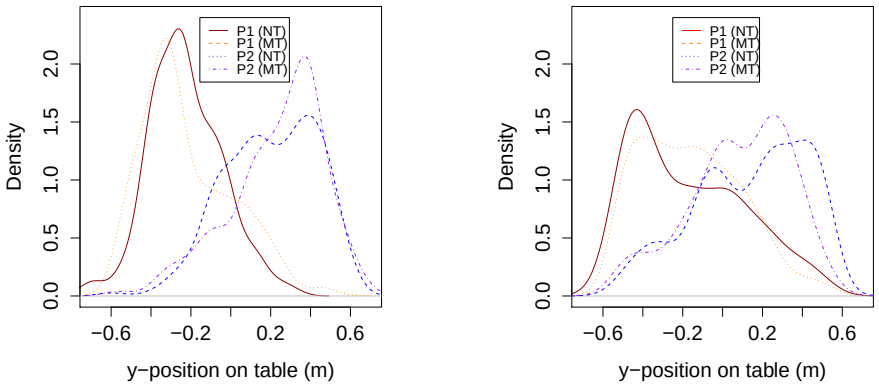

Fig. 16. Kernel densities of the pointing gestures for the fingertip positions (left) and the pointing target (right).

6 Discussion and Conclusion

For achieving robust and smooth human-robot interaction in shared space by anticipating upcoming gesture targets, the aim of this work was to study human gaze and pointing behavior in a face-to-face route planning scenario.

The experiments revealed that – when a pointing gesture is started – it is indeed possible to predict the target area by considering the fixations that happened directly before the gesture's onset. About one half of the target areas could be predicted when examining fixations 200 ms before the onset of the gestures. This provides a 500 ms (duration of common pointing gestures) advantage for interpretation, planning and reacting. With a probability of about 75%, it is possible to predict if a human hand just about starting to move will enter the robot's peripersonal space. We also examined the distribution of gesture targets: In the experiment, participants succeeded in partitioning their space equally, but for each participant, the percentage of pointing targets decreased relatively linear with distance. This suggests that the probability of conducting gestures decreases with the distance in face-to-face interactions. Reaching distant gesture targets was mostly handled by leaning forward. If the human partner starts leaning forward, it can safely be assumed that the target of the gesture will be distant (to the human). Thus, the gesture is likely to enter the robot's peripersonal space. In some cases, the pointing gesture was conducted from a higher position in some distance to the target.

Based on these findings, the robot's shared-space representation of Holthaus et al. [11] could now be extended by including such predictions. On the production side, the observed human skills can additionally be used to improve robot behavior. Fixating the target of a planned gesture shortly before conducting it could signal that the robot is going to occupy that area. Whether humans will use that information in this scenario remains to be tested.

Our method for automatically analyzing interactions facilitates follow-up studies with increased participant counts. Moreover, as our method can operate

in real-time, it can also be used in the human-robot interaction to make the robot aware of the interlocutor's gaze behavior. Future work will focus on creating models based on the presented findings to improve speed and robustness of the robot's spatial interaction.

7 Acknowledgment

This work has been partially supported by the German Research Foundation (DFG) in the Collaborative Research Center 673 Alignment in Communication. This paper is a preprint version of an article published by Springer-Verlag [26].

References

1. Abrams, R.A., Meyer, D.E., Kornblum, S.: Eye-hand coordination: Oculomotor control in rapid aimed limb movements. Journal of Experimental Psychology: Human Perception and Performance 16(2), 248–267 (1990)
2. Antonelli, M., Chinellato, E., del Pobil, A.P.: Implicit mapping of the peripersonal space of a humanoid robot. In: IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain. vol. 3, pp. 1–8. IEEE (2011)
3. Berti, A., Frassinetti, F.: When far becomes near: Remapping of space by tool use. Journal of Cognitive Neuroscience 12(3), 415–420 (2000)
4. Biguer, B., Jeannerod, M., Prablanc, C.: The coordination of eye, head, and arm movements during reaching at a single visual target. Experimental Brain Research 46(2), 301–304 (1982)
5. Breazeal, C., Kidd, C.D., Thomaz, A.L., Hoffman, G., Berlin, M.: Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In: International Conference on Intelligent Robots and Systems, 2005 (IROS 2005). pp. 708–713. IEEE (2005)
6. Carello, C., Grosofsky, A.: Visually perceiving what is reachable. Ecological Psychology 1(1), 27–54 (1989)
7. Farnè, A., Làdavas, E.: Auditory peripersonal space in humans. Journal of Cognitive Neuroscience 14(7), 1030–1043 (2002)
8. Green, A., Hüttenrauch, H.: Making a case for spatial prompting in human-robot communication. In: Language Resources and Evaluation (Workshop on Multimodal Corpora: From multimodal behavior theories to usable models) (2006)
9. Hadjidimitrakis, K., Breveglieri, R., Bosco, A., Fattori, P.: Three-dimensional eye position signals shape both peripersonal space and arm movement activity in the medial posterior parietal cortex. Frontiers in Integrative Neuroscience 6(37) (2012)
10. Holmes, N.P., Spence, C.: The body schema and the multisensory representation(s) of peripersonal space. Cognitive Processing 5(2), 94–105 (2004)
11. Holthaus, P., Pitsch, K., Wachsmuth, S.: How can i help? spatial attention strategies for a receptionist robot. International Journal of Social Robots 3, 383–393 (2011)
12. Holthaus, P., Wachsmuth, S.: Active peripersonal space for more intuitive hri. In: International Conference on Humanoid Robots. pp. 508–513. IEEE RAS, Osaka, Japan (2012)

13. Hüttenrauch, H., Eklundh, K., Green, A., Topp, E.: Investigating Spatial Relationships in Human-Robot Interaction. In: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 5052–5059. IEEE, Beijing (2006)
14. Kendon, A.: Conducting interaction: Patterns of behavior in focused encounters, vol. 7. CUP Archive (1990)
15. Làdavas, E.: Functional and dynamic properties of visual peripersonal space. Trends in Cognitive Sciences 6(1), 17–22 (2002)
16. Mark, L., Nemeth, K., Gardner, D.: Postural dynamics and the preferred critical boundary for visually guided reaching. Journal of Experimental Psychology 23(5), 1365–79 (1997)
17. Mumm, J., Mutlu, B.: Human-robot proxemics. In: Proceedings of the 6th international conference on Human-robot interaction - HRI '11. p. 331. ACM Press, New York, New York, USA (2011)
18. Neggers, S., Bekkering, H.: Ocular gaze is anchored to the target of an ongoing pointing movement. Journal of Neurophysiology 83(2), 639–651 (2000)
19. Nguyen, N., Wachsmuth, I.: From body space to interaction space-modeling spatial cooperation for virtual humans. In: 10th International Conference on Autonomous Agents and Multiagent Systems. pp. 1047–1054. International Foundation for Autonomous Agents and Multiagent Systems, Taipei, Taiwan (2011)
20. Nguyen, N., Wachsmuth, I.: A computational model of cooperative spatial behaviour for virtual humans, pp. 147–168. Representing Space in Cognition: Interrelations of behaviour, language, and formal models, Oxford University Press (2013)
21. Pfeiffer, T.: Understanding Multimodal Deixis with Gaze and Gesture in Conversational Interfaces. Shaker Verlag, Aachen, Germany (December 2011)
22. Pfeiffer, T.: Interaction between speech and gesture: strategies for pointing to distant objects. In: Efthimiou, E., Kouroupetroglou, G., Fotinea, S.E. (eds.) Gestures and Sign Language in Human-Computer Interaction and Embodied Communication, 9th International Gesture Workshop, GW 2011. pp. 238–249. No. 7206 in LNAI, Springer-Verlag GmbH, Athens, Greece (2012)
23. Pfeiffer, T., Renner, P.: Eyesee3d: A low-cost approach for analyzing mobile 3d eye tracking data using computer vision and augmented reality technology. In: Proceedings of the Symposium on Eye Tracking Research and Applications. pp. 195–202. ACM (2014)
24. Prablanc, C., Echallier, J., Komilis, E., Jeannerod, M.: Optimal response of eye and hand motor systems in pointing at a visual target. Biological Cybernetics 124, 113–124 (1979)
25. Previc, F.H.: Functional specialization in the lower and upper visual fields in humans: Its ecological origins and neurophysiological implications. Behavioral and Brain Sciences 13, 519–542 (1990)
26. Renner, P., Pfeiffer, T., Wachsmuth, I.: Spatial references with gaze and pointing in shared space of humans and robots. In: Freksa, C., Nebel, B., Hegarty, M., Barkowsky, T. (eds.) Spatial Cognition IX. Lecture Notes in Computer Science, vol. 8684, DOI: http://doi.org/10.1007/978-3-319-11215-2_9, pp. 121–136 (2014)
27. Rizzolatti, G., Scandolara, C., Matelli, M., Gentilucci, M.: Afferent properties of periarcuate neurons in macaque monkeys. Behavioural Brain Research 2(2), 147–163 (1981)
28. Staudte, M., Crocker, M.W.: Visual attention in spoken human-robot interaction. In: Proceedings of the 4th ACM/IEEE international conference on Human Robot Interaction - HRI '09. p. 77. ACM Press, New York, New York, USA (2009)

29. Velichkovsky, B., Sprenger, A., Pomplun, M.: Auf dem Weg zur Blickmaus: Die Beeinflussung der Fixationsdauer durch kognitive und kommunikative Aufgaben. In: Software-Ergonomie 97, pp. 317–327. Springer (1997)
30. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)