Evaluating Remote and Head-worn Eye Trackers in Multi-modal Speech-based HRI

Michael Barz
German Research Center for
Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3
Saarbrücken, Germany
michael.barz@dfki.de

Peter Poller
German Research Center for
Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3
Saarbrücken, Germany
peter.poller@dfki.de

Daniel Sonntag
German Research Center for
Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3
Saarbrücken, Germany
sonntag@dfki.de

ABSTRACT

Gaze is known to be a dominant modality for conveying spatial information, and it has been used for grounding in human-robot dialogues. In this work, we present the prototype of a gaze-supported multi-modal dialogue system that enhances two core tasks in human-robot collaboration: 1) our robot is able to learn new objects and their location from user instructions involving gaze, and 2) it can instruct the user to move objects and passively track this movement by interpreting the user's gaze. We performed a user study to investigate the impact of different eye trackers on user performance. In particular, we compare a head-worn device and an RGB-based remote eye tracker. Our results show that the head-mounted eye tracker outperforms the remote device in terms of *task completion time* and the required *number of utterances* due to its higher precision.

CCS Concepts

•Human-centered computing → Human computer interaction (HCI); •Computer systems organization → *Robotics;*

Keywords

human-centred computing; eye tracking; human-robot interaction; multi-modal interaction; machine learning

1. INTRODUCTION

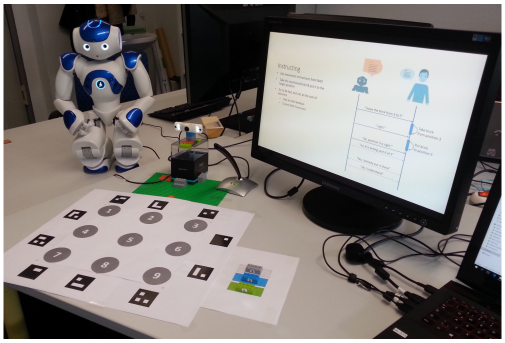

Figure 1: Setting of our user study.

Human gaze is involved in many processes in multi-modal speech-based interaction, such as in disambiguating speech, in joint attention during collaboration and in turn-taking [6]. For instance, there is a strong link between gaze behaviour and spoken language: speakers fixate elements ”less than a second before naming them” [2]. Further, the coordination of hand-movements involves vision, e.g., when ”directing the hand or object in the hand to a new location” [4]. To this end, gaze is ideal for incorporating non-verbal cues to human-robot dialogues, especially for complementing spatial information. We developed a multi-modal dialogue system that takes advantage of the temporal congruence between gaze and spoken deictic references to acquire and update spatial knowledge about objects. However, gaze estimation is erroneous [1] and the eye tracker's form-factor is important[1]. We conducted a user study for investigating the impact of different devices, a remote and a hear-worn eye tracker, on user performance. Figure 1 shows the setting of our study including a shared workspace (3×3 grid), the humanoid robot NAO and a display for calibrating the head-mounted eye tracker.

2. METHOD

Our dialogue system enables the humanoid robot NAO to learn new objects and their positions from the user (*Learning*), and second, can instruct the user to move objects while tracking this movement (*Instructing*). For *Learning*, we combine the user's speech and gaze to infer an objects position [2]. The link between gaze and hand movements facilitates tracking and updating position data [4] for *Instructing*. For realising the multi-modal dialogue interactions with NAO we used the Situation-Adaptive Dialogue Platform (SiAM-dp) [7].

We integrated two eye trackers, a head-worn Pupil eye tracker [3] as mobile device and a usual webcam with a modified version of *libfacetracker* [8] as state-of-the-art remote device. Both were adapted to report gaze in terms of normalised workspace coordinates, which could be mapped to a 3×3 grid (positions from 1 to 9) by computing the near-

[1] E.g., dementia day hospitals require non-obtrusive devices.

HRI '17 Companion March 06-09, 2017, Vienna, Austria

ACM ISBN 978-1-4503-4885-0/17/03.
DOI: http://dx.doi.org/10.1145/3029798.3038367

		M		SD	
Spatial Accuracy	Mobile	4.32°	42.94mm	1.17°	8.17mm
	Remote	4.69°	37.43mm	2°	8.44mm
Spatial Precision	Mobile	.37°	3.83mm	.28°	3.37mm
	Remote	2.36°	20.08mm	.78°	3.15mm

Table 1: Mean and SD of spatial accuracy and spatial precision in degrees of visual angle and mm.

est neighbour (Euclidean distance). For the mobile device, we used the built-in screen-based calibration algorithm and the marker-based surface detection for mapping gaze to the workspace. Real-time fixation detection is not supported, hence we apply our own dispersion based method. For the remote device, we extended *libfacetracker* with a head pose estimation similar to [5] to receive 3D gaze estimates. These could be intersected with our workspace ($z = 0$ plane). Further, we developed a polynomial calibration feature to automatically configure the camera position and to cope for individual differences. Calibration is required once per user.

3. EVALUATION AND RESULTS

To evaluate our dialogue system we conducted a within-subject user study (10 participants). We tested how users perform with two different eye trackers, a *Mobile* and a *Remote* one, in two tasks, *Learning* and *Instructing*. For *Learning* we asked the user to teach three objects (LEGO® bricks) to NAO by putting them on the grid between them, one by one, stating, e.g., *"This is a brick"*. For *Instructing* all bricks were randomly distributed on the 3×3 grid and the user had to follow three instructions of NAO, e.g., *"Move the brick from position 3 to 5"*. Besides, we asked users to perform an *accuracy and precision test* with each device (they had to fixate pre-defined workspace targets). For the remote device, these samples were used for calibration.

In a first step, we analysed the spatial accuracy and spatial precision of both eye trackers averaged over all sampling targets (see Table 1). The differences in mean were significant for spatial precision ($t(9) = 9.25, p < .001$), but not for spatial accuracy ($t(9) = -1.43, p = .186$). In a second step, we analysed the *task completion times* and the required *number of utterances* for both eye trackers, separate for each task. For both measures, the mobile device achieved significantly better results (see Figure 2). On average, it took participants 4.4s and 5.28s less to complete a task using the mobile device for *Learning* ($t(9) = 3.42, p = .008$) and *Instructing* ($t(9) = 3.57, p = .006$), respectively. Likewise, the required *number of utterances* was decreased by 0.92 ($t(9) = 4.12, p = .003$) and 0.96 ($t(9) = 3.59, p = .006$), respectively. We used the two-tailed paired samples t-test.

4. DISCUSSION AND CONCLUSION

Our evaluation showed that the mobile eye tracker is significantly more precise than the remote device. However, both devices have nearly the same spatial accuracy concerning the shared workspace of our study. Reasons for the inaccurate results of the mobile device, which we expected to better in this regard, are probably due to parallax error. Nevertheless, the mobile eye tracker outperformed the remote device in terms of *task completion time* and required *number of utterances* (all utterances of the user and NAO). This leads us to the conclusion that high spatial precision is essential for our human-robot collaboration and thus the mobile eye tracker is better suited for our scenario.

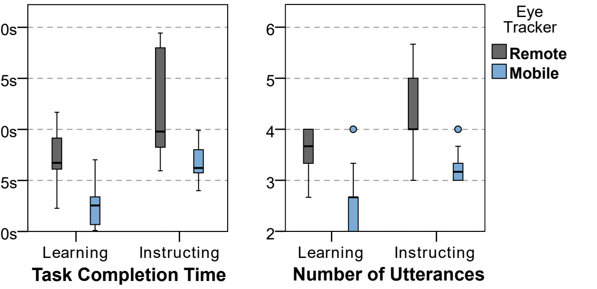

Figure 2: Dependent measures for each condition.

In future work, we aim to include further eye tracking devices for investigating the impact of additional features such as spatial accuracy. More sophisticated calibration and fixation detection techniques could increase the performance of our dialogue system, as well. In addition, we want to enhance the gaze mapping process to the 3D environment, to become independent of the grid-like interaction space. A limitation of our system is the timing of fusing gaze and deictic references, which could be solved by incorporating the user's hand movements or learning individual delays.

5. ACKNOWLEDGMENTS

We thank our project partners in CPS for smart factories (EIT Digital), http://dfki.de/smartfactories/.

6. REFERENCES

[1] M. Barz, F. Daiber, and A. Bulling. Prediction of Gaze Estimation Error for Error-Aware Gaze-Based Interfaces. ETRA '16, pages 275–278. ACM Press, 2016.

[2] Z. M. Griffin and K. Bock. What the Eyes Say About Speaking. *Psychological Science*, 11(4):274–279, 2000.

[3] M. Kassner, W. Patera, and A. Bulling. Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction. UbiComp '14 Adjunct, pages 1151–1160, New York, NY, USA, 2014. ACM.

[4] M. Land, N. Mennie, and J. Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28(11):1311–1328, 1999.

[5] S. Lemaignan, F. Garcia, A. Jacq, and P. Dillenbourg. From Real-time Attention Assessment to "With-me-ness" in Human-Robot Interaction. HRI '16, pages 157–164. IEEE Press, 2016.

[6] G. Mehlmann, M. Häring, K. Janowski, T. Baur, P. Gebhard, and E. André. Exploring a Model of Gaze for Grounding in Multimodal HRI. ICMI '14, pages 247–254. ACM Press, 2014.

[7] R. Neßelrath. *SiAM-dp : an open development platform for massively multimodal dialogue systems in cyber-physical environments*. PhD thesis, Universität des Saarlandes, 2015.

[8] Z. Tosér, R. A. Rill, K. B. Faragó, L. A. Jeni, and A. Lörincz. Personalization of gaze direction estimation with deep learning. In *KI 2016: Advances in Artificial Intelligence*, pages 200–207, 2016.